University of Southampton

Faculty of Engineering and Physical Sciences

Electronics and Computer Science

Reconfiguring Open Data for Open Innovation

by

Johanna Catherine Walker

ORCID ID 0000-0002-5498-8670

Thesis for the degree of Doctor of Philosophy in Web Science

February 2020

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
Electronics and Computer Science
Web and Internet Science

Thesis for the degree of PhD in Web Science

Reconfiguring Open Data for Open Innovation

by

Johanna Catherine Walker

This thesis explores the praxis of open data in a public sector open innovation environment. It seeks to understand how the processes of the use of open data are operating in this context and locates the conditions that may cause it to diverge from theory. To do this, it first undertakes a literature review on open data and open innovation, which brings both of these 'open' practices together and establishes a model of open data for open innovation.

Taking a positive, rather than normative approach, the thesis develops a case study of four mid-sized cities engaging in open innovation to create solutions to municipal problems over a period of three years. The data sources for this case study are the documents created in and for the open innovation projects, and a group interview of representatives of the cities. Via this case study, the thesis makes the key assertion that the open data ideal is operating under both regulatory and resource constraint, which obstruct, rather than support, the attempt to capture value from data.

An integrative literature review is used to investigate how (other) forms of data sharing can inform the development of a model that reflects praxis. By comparing key aspects of the open data model with insights derived from data sharing, this thesis demonstrates how the open data model can be reconstructed to ensure legal and associated compliance that will protect and promote innovation with data.

The outcome of this research is a new lens with which to view the use of open data. Furthermore, it offers concrete suggestions for enacting changes to open data processes that will enable greater productivity of data-driven innovation. As such, it has vital policy and practice implications for open data and data sharing.

Table of Contents

Table of Tables

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Johanna Catherine Walker

Title of thesis: Reconfiguring Open Data for Open Innovation

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

This work was done wholly or mainly while in candidature for a research degree at this University;
Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
Where I have consulted the published work of others, this is always clearly attributed;
Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
I have acknowledged all main sources of help;
Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
Parts of this work have been published as:- Walker, J. Simperl E. and Carr, L., (2019) A framework for data sharing for open innovation. *18th Open and User Innovation Conference*, *(OUI19)* July 8- 10th, 2019, Utrecht University, the Netherlands

Signature: ..... ……Date: 13 November 2020

Acknowledgements

For his help and support throughout my PhD, I would like to thank my main supervisor, Professor Les Carr. His conviction that completing this thesis was indeed possible was a major motivator when completion seemed, all too often, impossible. For this, his insights and his general air of being a stable ship on the tumultuous sea of the PhD, I am most grateful. I would also like to thank my secondary supervisor, Dr Tom Wainwright (now of Royal Holloway).

My family and friends have been remarkably supportive over an extended period of time, for which they deserve sincere thanks. I would also like to thank the wider open data community, especially at the Open Data Institute and Open Data for Development, who have provided such opportunities for research and dissemination.

The research in this thesis would not have been possible without the generous and whole-hearted cooperation of the city partners of the Interreg2Seas project Smart Cities Innovation Framework Implementation (SCIFI). They are some of the most interesting, committed and open people I have ever had the pleasure to work with, and I thank them and the other members of the SCIFI consortium who provided me with the data and insights necessary for this thesis. I would also like to thank Professor Elena Simperl, who introduced me to this and many other open data and data sharing research opportunities over the past three years.

Thanks are also due to my colleagues with whom I have researched and published over the past 6 years, both at the University of Southampton, in the wider Digital Economy Network and beyond, for encouraging me to think in a way that is truly interdisciplinary.

Finally, I must thank the Web Science Centre for Doctoral Training and Web and Internet Science research group which has been my academic home during this process, along with my funders, the EPSRC Digital Economy Programme. I would like to thank my lecturers, professors and everyone within WAIS, Electronics and Computer Science and the wider University who together create the entire PhD experience, of which the thesis is only a part.

Definitions and Abbreviations

Abbreviations

SME: **S**mall and **M**edium **E**nterprises
DPbD: **D**ata **P**rotection **b**y **D**esign
FOI: **F**reedom of **I**nformation (Act/Request)
GDPR: **G**eneral **D**ata **P**rotection **R**egulation
OKF: **O**pen **K**nowledge **F**oundation
PSI: **P**ublic **S**ector **I**nformation
API: **A**pplication **P**rogramme **I**nterface
PRISMA: **P**referred **R**eporting **I**tems for **S**ystematic Review and **M**eta-**A**nalyses

Definitions

Absorptive Capacity:	An organisation's ability to identify, assimilate, transform, and use external knowledge
Business Model:	How an organisation will create revenue while serving the customer base
Civic Accelerator:	A programme to create solutions to public sector challenges by providing mentoring, coaching and financial investment to support start ups and SMEs to work on solutions
Data Collaborative:	Participants from different sectors exchange their data to create public value
Data Commons:	Platforms that co-locate data, infrastructure, and tools and services to create a resource for managing, analyzing and sharing data with a community
Data Marketplace:	Platform where entities buy and sell data
Data Steward:	Role within an organization responsible for utilizing an organization's data governance processes
Data Trust:	Legal structure that provides independent stewardship of data for the benefit of a group of organisations or people
Digital Innovation Contest:	An event in which third party developers participate to develop and implement a service prototype for a particular aim, based on (open) data
GDPR:	General Data Protection Regulation, European regulation to ensure the free flow of data while preserving privacy
Hackathon:	A particular type of digital innovation contest, usually time-bounded and non-virtual
Internet of Things:	Network of physical objects embedded with sensors and other technologies for the purpose of exchanging data with other devices and systems over the Internet
Open Data:	Data that is legally available for anyone to use for any purpose
Open Government Data:	PSI that is licenced as open data
Open Innovation:	A model of innovation proposed by Henry Chesbrough, focusing on flows of ideas across organisational boundaries
Public Sector Information:	Data collected or generated by the Public Sector. This may be publicly available but is only open data if accompanied by the appropriate licence
Quadruple Helix:	A framework describing interactions between academia, government, business and citizens in a knowledge economy
Shared Data:	Direct access provided to data to named parties for a specific purpose
Smart City:	A programme for developing liveable, sustainable cities with the intersection of data, technology and infrastructure

Chapter 1 Introduction

1.1 Open Data, Open Innovation and the Public Sector

The incentives for opening data are often framed as wildly economically valuable. "*Data is the new oil" (*Dietrich et al 2009) and "*open data could help unlock $3tn of value*" (Manyika et al, 2013) are often-repeated, attention-catching statements that have been used by advocates to promote the value of open data. It is promoted as possessing significant potential to stimulate innovation. It is used to create new products and services by entrepreneurs, and published by corporations and organisations to encourage the creation of new ideas outside the firm's boundaries. Newer companies such as CityMapper use datasets such as map data and public transport timetables to create their applications. More traditional companies, such as Yorkshire Water, use it to increase an understanding of industry analytics.

Measurement of impact and outcome of open data are mostly expensive and subject to issues with proving causality (Lammerhirt and Brandusescu, 2019). However, statistical and narrative assessments suggest the promise has not quite been achieved. Narrative indications suggest that the innovation aims of open data are still very much emergent. The GovLab's Open Data Impact bank of case studies shows 6 case studies under the heading of 'creating opportunity' (for business and innovation). However, of these, only two are actually businesses - the other four are stories of release of data that are hoped to have potential for commercial innovation. One of the few longitudinal studies of both publishing and outcomes, the Open Data Barometer, notes that, even though they believe governments are prioritising the release of innovation-linked data, fewer relevant datasets were opened globally in 2016 than in previous years, and they suggest that what is published is underused. A report from NESTA suggests that innovation with Open Data is nowhere near expected levels (Rubenstein, Cowls and Cath, 2016). In the UK, and further afield, the question of how best to engage with users and potential users is still open.

Government is important in the open data environment for a variety of reasons. By multiple orders of magnitude, the public sector is the largest publisher of open data, and 'open government data' is often synonymous with open data in practical terms. In the UK and much of Europe, it is government policy that determines the open data narrative. The focus of policy as investment in portals and innovation via the provision of support for start-ups, means that these are the dominant areas of interest.

The limited success around open data impact is, in the main, ascribed to failings associated with these programmes – particularly portals. Improving usage, in this context, is a case of improving current provision. Brandusescu, Iglesias and Robinson (2017), for instance, believe that usage will be increased if the *"user-friendliness*" of portals is increased. However, Janssen, Charalabidis and

Zuiderwijk (2012) contend that much impetus surrounding the release of data subscribes to the 'if we build it, they will come' fallacy, regardless of how well 'it' is built. Consequently, there is a space for investigating demand-side, as well as supply-side, issues of open data.

That space is defined by two fundamental elements: data (the potential value of which is acknowledged above), and openness. The proliferation of activities characterised as 'open' since the beginning of the century has led to a certain amount of ambiguity in the term (Longshore Smith and Seward, 2017). Possibly building on the success of the free, libre and open source software (FLOSS) movement, this includes concepts such as open government, open democracy, open access, open science, open knowledge and open Internet, amongst others.

One such activity is open innovation. In open innovation, organisations rely on sources of knowledge that are located outside the boundaries of the organisation. The mechanisms for identifying, absorbing and executing on the external knowledge vary. While some argue that open innovation is merely a semantic, rather than genuinely novel, concept, it is a natural ground for exploring theories of how innovation can be created with open data.

Government has been slower in adopting open innovation. There are particular rules and regulations surrounding government organisations that potentially limit their ability to innovate (Kankanhalli, Zuiderwijk and Kumar Tayi, 2017). However, there are emerging policy and service challenges that the public sector needs to address that are accelerating a need for external input.

The vast majority of open innovation initiatives in the public sector have involved initially opening up public sector information (PSI), then a subset of the same as open data. These efforts have mainly focused on processing data and presenting this to citizens and businesses (Kankanhalli, Zuiderwijk and Kumar Tayi, 2017). Local governments use wireless sensor networks to efficiently manage their cities and improve public welfare. This data has become associated with the rise of the 'smart city' paradigm. Information from sensor networks in open data platforms is a new area in which to "*foment*" competition, economic growth and citizen welfare (Domingo et al, 2013). Businesses and developers (seen as potential future entrepreneurs) are the most explicitly referenced users in EU open data policy documents (Lassinantti, 2014). Their role is to exploit the data for economic value and job creation, via engaging in (open) innovation with open data.

My research questions and objectives are situated in this relationship between open data, open innovation and the public sector.

1.2 Research Questions and Objectives

This research interrogates the praxis of open data for open innovation in the public sector – how city authorities are acquiring new solutions to civic challenges created by small companies with the city's open data. It seeks to understand how the processes of the use of open data for open

innovation are operating in reality and locates the conditions that may cause it to diverge from theory. It investigates this space, and derives a framework of open data that reflects productive yet legitimate use.

This thesis comprises three research questions.

RQ1: What are the key components of a framework of open data for open innovation?

RQ2: How does the use of open data in open innovation in practice vary from the framework defined in RQ1?

RQ3: How can comparison of other types of public and private data sharing arrangements inform the framework defined in RQ 1 so it more accurately reflects open data for open innovation as found in practice?

General objective: To understand how open data is being used for open innovation in practice, if this reflects the use presented in the literature, and how any differences affect the theoretical model.

RO1: To outline the current research on open data, open innovation and open innovation with open data;

RO2: To assess the knowledge, attitude and practice of open innovation users towards open data;

RO3: To identify practices within the data sharing literature that reflect, and can support, how open data is used for open innovation in practice.

1.3 Contribution

The contribution of this thesis is threefold:

It compiles and presents a novel literature review on open innovation with open data - previous literature reviews have focused on these dimensions separately;

It demonstrates how the 'rules' of open data are eroded in the attempt to capture value from data due to regulatory and resource constraints, and how these can be reconstructed to ensure legal and associated compliance that will protect and promote innovation with data;

It presents research on open innovation in a government context, an area where more research is needed (West and Bogers, 2017; Kankanhalli, Zuiderwijk and Kumar Tayi, 2017). It also focuses the research on open innovation in a consortium, which has been under-researched outside the computing and communications industries according to West and Bogers (2017).

The outcome of this is a new lens with which to view the lack of impact of open data, and more importantly, concrete suggestions for enacting changes to open data structures that will enable greater productivity of data-driven innovation.

1.4 Scope of the Thesis

This thesis examines the use of open data in open innovation in a public sector setting. This has naturally arisen as the vast majority of open data is open government data. This means that the public sector is therefore part, even if passively, of the majority of open innovation with open data. Consequently, much of the existing research features the public sector in a greater or lesser role of open innovation with open data. With the advent of a focus on smart cities, public sector open data and open innovation is likely to continue as the dominant form, and this, therefore, makes a pragmatic choice of focus.

1.5 Outline of the Thesis

This introduction has provided the context from which this research originated. In Chapters 2, 3 and 4, I conduct a literature review of related research that achieves two aims. Firstly, it contextualises and justifies my research questions and approach. It sets the theoretical framework of open data, open innovation and open innovation with open data. From this, it naturally focuses on the public sector. Secondly, this literature review allows the evaluation of the state of knowledge on open data and open innovation, addressing Research Question 1. The production of knowledge continues to accelerate, and becomes increasingly both interdisciplinary and fragmented, as new disciplines emerge. Therefore, the literature review as method is "*more relevant than ever*" (Snyder, 2019).

In these chapters, the literature is reviewed on a topic-centric, rather than publication-centric, basis. This allows for criticality and also brings the researcher's perspective on the subject to the fore (Khoo, Na and Jaidka, 2011). An approach of this nature privileges the development of argument and interpretation of the material by the researcher. The locating and collection of the material follows a hermeneutic process (Boell and Cecez-Kecmanovic, 2014). Instead of formulating a priori terms and boundaries for inclusion and exclusion of literature this iterative approach locates and critically assesses material and develops arguments, leading to further search and review. In this way, the engagement with development of understanding of material leads to further questions and search. In a relatively new area such as open data, the end of the process comes as both novelty of argument and citations are diminished. The final output of these chapters is a framework of open data for open innovation, which is used to establish key themes that will enable a deductive approach for the main body of the research. A framework is a 'supporting structure', and this framework represents the requirements for open data that need to be in place for open innovation to take place.

Chapter 5 presents the Smart Cities Innovation Framework Implementation project, which is the case study for research question 2. This is an open innovation programme wherein 4 cities in

northern Europe seek to harness the power of the market to create solutions to public sector challenges, using their open data and other open innovation instruments, namely a digital innovation contest and a civic accelerator. The format of the open innovation programme is outlined and the cities introduced.

In Chapter 6 the methodology for addressing Research Questions 2 and 3 is outlined. This is informed by previous open data research, where the qualitative approach and case studies have been suggested as suitable (Davies, Perini & Alonso, 2013).

The overall research strategy for Research Question 2 is a case study of the Smart City Innovation Framework Implementation project. This involves four European cities which are opening governmental data for open innovation with six private companies. Triangulation is adopted and three research strands are designed in order to thoroughly investigate the case study and provide insights into open data use.

The first research strand is textual document analysis. As innovation is first and foremost a process, this method allows interpretation of sources produced over a series of years and for a range of purposes, to build up a wide picture of multiple aspects of the related activities. The second research strand is an analysis of the data utilised in the open innovation activities. These two approaches ensure that source data regarding how open data was actually used in practice was captured, rather than recording the cities' interpretations of their use of open data. The results of these two methods are then triangulated with a group interview with the representatives of the cities participating in the project. This seeks to achieve two things: firstly, gain confirmation that the results of the data analysis are accurate, and secondly to acquire the reactions and insights of the city representatives to the overall results of how they used open data in practice for open innovation. In Chapter 7 I analyse the results of these three research strands through content and thematic analysis, and in Chapter 8 I discuss the implications of the findings.

These findings create the basis for Chapter 9. In this I address Research Question 3, which investigates how open data use in practice can be remediated or adapted, by developing a model for data sharing for open innovation through an integrative review of the data sharing literature. An integrative review is undertaken with the aim of assessing, critiquing and synthesizing the literature in a way that enables new perspectives to emerge (Torraco, 2005). It is particularly suitable for nascent, embryonic topics. In Chapter 10, a discussion is then presented regarding meaning and impact, before the final conclusions are made and an adjusted model for open innovation with open government data is proposed.

In the final Chapter, I reiterate my findings and present my conclusions, and again summarise the key points of the model. Lastly, I discuss directions for future work.

Chapter 2 Literature Review: Open Data

2.1 A Brief History of Open Data

Open data, which can most simply be defined as data that can be freely reused by anyone for any purpose, has multiple roots. On the one hand, it has been driven by the increased ability for sharing of documents and data delivered by the World Wide Web, and as such has its base in movements such as free and open source software. On the other, it has been powered by a need for democracy, knowledge and transparency, with beginnings in the Freedom of Information movement and driven by groups such as the Open Knowledge Foundation (OKF). Although not exclusively a public sector movement, it has a close relationship to open government and civic technology groups. As Harrison, Pardo and Cook (2012) note, "*At the heart of the open government ecosystem is the assumption that government possesses information that users want and will use.*" In terms of the historical development of open data, Open Knowledge's annual report for 2004-2005 makes no mention of public sector data or open data and focuses on the open knowledge definition. In 2005, the open definition – defining 'open' in relation to both data and content was first developed. By the time of the subsequent annual report, 2006-2007, Open Knowledge had developed the Comprehensive Knowledge Archive Network (CKAN) and had reached, "*General agreement …on the importance of access to public data, in a raw form, with accompanying identifiers*." (OKF, 2007). On his first day in office in 2009, US President Obama signed the Memorandum on Transparency and Open Government asking government agencies to release their data to make it open and available to the public. The International Open Data Charter Principles were launched in 2015, led by a coalition of open data interest groups. Subsequently, the Open Data Charter has been adopted by 73 national and local governments. As of 2017, all EU countries had Open Data policies which encouraged re-use, although these vary in their development (Carrera et al, 2017).

There are philosophical, legal and economic arguments for open government data. Philosophically, government data belongs to the citizens, in the service of whom it is collected, and whom it is about. Legally, Freedom of Information laws mean that citizens can access (somewhat arduously) information on government activities as a fundamental right.

In 2008, the economist Rufus Pollock argued that the efficient price for public sector information for users was marginal cost – effectively zero with digital distribution. By 2012, Chris Yiu made the policy case for all non-personal public sector information to be made available for free. Estimating the net loss to the UK government of data reselling at £50m per annum, Yiu argued that the benefits of opening all publicly held data would outweigh the loss of revenue many times over. The following year, Manganika et al (2013) calculated that open data was worth up to USD3trillion

to the global economy in efficiencies, innovations and consumer value. Governments that open up data may gain financially through two channels: increased employment leading to lower unemployment subsidies and higher tax revenues, and higher indirect tax revenue from related products and services. Additionally, the public sector benefits from significant efficiency gains and reduced transaction costs. The European Data Portal estimates that 25,000 jobs will be created by Open Data in 2020, and more than 30 million euros of public administration savings will be made in 11 countries (Berends et al, 2017). A timeline of key events in open data is shown on page 10.

In 2013, Open Knowledge launched the Open Data Index, which identified 'key' data sets through process of "*discussion and consultation*" driven by the Open Government Data Working Party who identified three benefits, transparency, participatory governance and social and commercial innovation (Pollock, 2013). However, this has so far failed to translate into the anticipated impact. Worthy (2013), asking *"Where are the armchair auditors?"* suggests that a "*missing link*" with regard to what should be done with information once acquired and analysed has reduced the impact. Understanding and fostering data use, rather than merely providing access, is critical to create value from open government data (Ubaldi, 2013).

Table 1 Timeline of Open Data

>>Pre-open>>			>>Preparation>>			>>Experimentation>>		
2003	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Public Sector Information Directive 2003/98/EC	Open Knowledge Foundation (OKF) report has no mention of open data; TheyWorkForYou launched with scraped Hansard data	Open Definition launched	Measuring EU Public Sector Information Resources study estimates potential PSI reuse value at €27bn	CKAN catalogue platform launched; 8 principles of open government data launched (opengovdata.org	Apps for Democracy launched to encourage reuse of the Washington DC data catalogue	Memorandum on Transparency and Open Government; data.gov launched; aporta.es (now data.gob.es) launched	Tim Berners Lee proposes 5 stars of Linked Open Data metric; Sir Nigel Shadbolt proposes pan-European data portal	OKF launch pan-European Open Data Challenge digital innovation contest

>>Evangelisation>>			>>Institutionalisation>>					?
2012	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
Open Data Institute founded	Amendment of Directive 2013/37/EU (included new bodies, limited fees, required machine readability; first edition of the Open Data Barometer	Open Data Now, first book on open data, is published; OpenData500, tracking companies using open data, launched	1st International Open Data Conference held in Ottawa; Open Data Charter launched; European Data Portal launched	Numerous open data portals, activities and indices in existence.	All EU member states have Open Data policies encouraging reuse: General Data Protection Regulation comes into effect	Impact Assessment Support Study for the Revision of the PSI Directive estimates baseline value of open data market as €52bn	Public Sector Information Directive replaced by 'Open Data Directive' (EU) 2019/1024	

This literature review takes a linear thematic approach to the open data literature, looking at both supply and demand, and moving from the decision to release to innovative use.

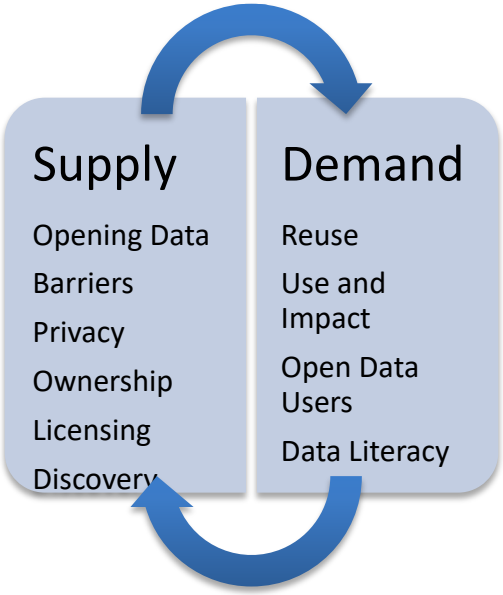

Figure 1 Summary of Open Data Topic in Literature Review

2.2 The Open Definition

In its simplest form, the open definition reads, "*Open data [and content] that can be freely used, modified and shared by anyone for any purpose*" (OKF, 2005). The longer definition notes the data must be in the public domain or appropriately licensed, no extra terms must be required for use, which must not be limited, there must be no discrimination against any person or group who wishes to use it and it should be downloadable from the internet at cost.

The 6 principles of the International Open Data Charter (IODC, 2015) state that open data should be:

- Open by Default
- Timely and Comprehensive
- Accessible and Useable
- Comparable and Interoperable
- For Improved Governance and Citizen Engagement, and,
- For Inclusive Development and Innovation.

Longshore Smith and Seward (2017) suggest that individual definitions of openness associated with each type of digital artefact (such as open data or open knowledge) muddy the concept of openness with the plurality of meanings. They characterise the meaning of openness across a

range of more than 50 contexts and artefacts as, 'you don't have to pay' and 'anyone can participate'. However, they also note that both of these concepts are only theoretically achievable. Even if there is no upfront cost for the data itself, accessing open data requires a computer and broadband, both of which have costs associated. Participation in the digital sphere may have cultural or political barriers, and requires a certain level of numeracy and literacy.

2.3 Opening Data

Early work in this area focused extensively on access to open data - what should be opened, licenses, quality control, personal data, data validation and authentication, funding and dissemination (Arzberger et al 2004, Cabinet Office, 2011). Much work was also framed in terms of the barriers to open data (Zuiderwijk et al, 2012, Janssen et al, 2012, Martin et al, 2013). At the core of this is the question, how do data owners decide what data to open?

Some data is mandated for opening at national and supra national levels. All European countries now have open data policies. In an attempt to make data available for economic growth, the EU published the Public Sector Information Directive 2003/98/EC, which established a set of minimum rules governing the re-use and the practical arrangements for facilitating re-use of existing documents. This has since been superceded by the Open Data Directive, 2019/1024. However, publishing strategies to fulfil these directives must still be found.

In a comparison of approaches taken by two Czech agencies, Kucera and Chlapek (2013) identify what they characterize as the 'top down' and 'bottom up' strategies for data opening. They conclude that both methods have benefits: bottom up allows data publishers to start quickly and learn from experience. They suggest that the analysis stage of the top down approach allows data owners to better know what data they possess and its characteristics. From an empirical point of view they note that the agency adopting the 'bottom up' approach appears to have greater impact, but they refrain from assigning this to the publishing method itself.

The UK government consulted with a large number of bodies and individuals when initially deciding which data to open and how to present it (Cabinet Office, 2011). There was common agreement on the need to develop a data inventory but less on how to deliver it (via centralised portal or sectorally). User experience and prioritisation were key parts of release, with user experience considered paramount. Factors determining prioritisation were limited to existing demand as evidenced by Freedom of Information requests. Another factor considered was the time and cost required for preparing the datasets.

In Germany, the municipal government of Berlin asked citizens which data sets they were interested in releasing. "*Instead of deciding what open data to focus on by replicating the focus taken in other cities, we asked people in Berlin about their dataset priorities in an (anonymous)*

online vote: a kind of crowd-sourcing," (Both, 2012). Citizens could select up to three content areas from predefined categories partially based on the structures of other cities' data catalogues. Hivon and Titah (2015) argue that Open Data websites should not be top down only, and posit that citizen participation can improve this.

Some national and sub-national governments, such as the UK, London and Montreal, have request systems alongside their portals and catalogues, although this is a very limited approach (Granickas, 2014). It is also difficult to find evidence of governments listening to feedback from users. During the period from September 2012 to May 2015, the UK government opened up 5 datasets in response to requests on data.gov.uk, during which period it received over 800 requests. Shekhar and Canares (2016) note that the governments of Buenos Aires and Montevideo, "*have not used technology to [....] establish feedback loops that would be durable or could connect with hard to reach communities*." The central government of the UK was, until recently, one of the few examples of governments that met regularly with independent advisory panels for this purpose. However, as of May 2015, none of these bodies have retained their mandates. Freedom of Information requests are another form of time-specific, often citizen-specific data request, and one benefit of open data is posited to be reducing the time spent on meeting these requests. However, there is no evidence so far that opening data has substantially reduced FOI requests.

Lee and Kwak (2011) point to the Pareto Principle, which suggests that governments should publish 20% of their data, which will be used around 80% of the time. However, this approach does not fit with innovation, fails to take the 'long tail' facilitated by the Web into account and does not help decide what to open.

2.4 Barriers to Open Data

The socio-technical risks, challenges and barriers to opening data include the cost, other resource capacity, concerns about quality, concerns about privacy and lack of a data culture within data holding organisations (Zuiderwijk et al, 2012; Barry and Bannister, 2014; Conradie and Choenni, 2012; Janssen et al ,2012; Martin et al 2013 and Walker, 2014). Kucera et al (2015) define the following categories of barriers: political and social; economic; organisational, legal and technical. There are a number of different types of political risk. The first is that data release and use of data can have unintended negative consequences even when the intended consequence is positive (Krishnamurthy and Awazu, 2016). Dulong du Rosnay and Janssen (2014), note that public bodies fear losing control over their data, but also have concern for any liability they might incur, should their data be erroneous, or misused. Even correct data may carry political risk: for instance, sensors can reveal how cities are becoming smart unevenly - with potential consequences (van Zoonen, 2016). Authenticity is a particular concern, especially for registers or other canonical

sources of data where a user needs to be assured that no unauthorised alteration has been made to the data (and may also need a guarantee of authenticity of some kind). Two such approaches use Merkle trees and the more comprehensive blockchain technology (Harrison, 2016; Potter, 2015). Government privatisation strategies can also disrupt the provision of open data, where, for instance, the data is no longer the government's to provide, but instead is under the control of private organisations.

Economic challenges often cluster around the cost of implementing and sustaining open data publishing programmes. The formatting of some government departments (for instance, the UK's Driver and Vehicle Standards Agency, the Meteorological Office and Companies House) as 'trading funds', which are required to self-finance, creates tension with data opening strategies (Martin et al, 2013). Rogawski, Verhuulst and Young (2016) present an example of how the business model of the UK Ordnance Survey conflicts with opening data. The article emphasises that making data freely available not only reduces the opportunity to create a direct income stream from that data, it also imposes a burden of data maintenance upon the publisher, which must somehow be funded. On the other hand, private companies may contest the publication of data that they feel unfairly undermines data that underpins their business. When the Dutch Ministry of Infrastructure and the Environment considered opening up its road database, some private sector mapping companies challenged this, as they felt offering a ready-made road database to their competition was unfair (Dulong du Rosnay and Janssen, 2014). In the UK, an attempt to replicate the Postcode Address File, the database of all postcodes in the UK, which had been sold to a private company, ran into insurmountable legal barriers.

Assuring citizens' privacy and reducing the likelihood of open data adding to existing levels of risk of privacy harms are also key sources of tension. Gomer, O'Hara and Simperl (2016) call privacy an, *"inevitable concern about open data."* The authors designate the relevant sphere of privacy in relation to open data as 'informational privacy'. They note that this type of privacy can be equally breached by details being excluded from datasets as being included, as this leads to "*impertinent enquiry and inference*". However, they note that in some cases it is desirable to publish personal information, such as contact names. Privacy-oriented activities, such as anonymisation, can be disrupted by open data usage activities, such as linking datasets. The authors also note that the fear of reputational damage, incurred by the revealing of a potential for actual harm through a privacy violation, is a concern for both publishers and consumers. They promote a two-pronged approach for managing privacy that involves consent management and stakeholder dialogue.

2.5 Privacy

It is a key tenet of open data that it is never personal data, thus protecting the privacy of individuals. In Europe, the General Data Protection Regulation (GDPR) was introduced on May 25, 2018, replacing the 1995 data protection directive. Article 1.1 states that it "*lays down rules relating to the protection of natural persons with regard to the processing of personal data and rules relating to the free movement of personal data."* (Voight and von dem Bussche, 2017). These are the key remits of the GDPR: to protect the rights of natural persons when it comes to data, and to support the free flow of data between Member States.

GDPR therefore only applies to personal data. Personal data is, "*information that relates to an identified or identifiable individual*" according to the UK Information Commissioner's Office, and could be immediately identifiable, such as a name, or identifiable by cross-checking, such as an IP address. Open data is 'never personal data'; however, this is an over-simplification of the situation. There are several reasons why this is so.

GDPR has complicated - or possibly clarified - an ongoing discussion around the boundaries of Public Sector Information that is publicly available, and open data. In this context, it is important to note that any information relating to an identified or identifiable natural living person, be it publicly available or not, constitutes personal data. There is such a thing as publicly available personal data, such as the name of a CEO of a company, or owner of a specific area of land. However, the fact that data has been made publicly available does not mean the GDPR does not hold. Issues around compliance when working with data that can be openly published but not necessarily processed are not new to the GDPR, just perhaps more visible (Hanecak, 2017).

The reuse of personal data made publicly available thus remains subject in principle to the relevant data protection law. For example, in the OpenActive project, where people might wish to choose their exercise classes based on specific coaches and activity leaders. Although the names of class leaders are freely available online, it still constitutes personal data, and therefore can only be used with consent - which in this particular example, is not impossible to gain (Dodds, 2018).

Secondly, it is possible for personal data to appear somewhat impersonal. There is a far broader range of data that can be construed as personal - in that someone may be directly identified from it - than may at first be imagined. It covers any kind of data that 'relates to' an individual and causes them to be identified. This may even be data such as information on journey routes and times, if they enable an individual to be identified (Young, 2018).

Data can be made personal due to the purpose of use - i.e. it could become personal data (Stalla-Bourdillon et al, 2020). If the purpose of the use of the data is to "*evaluate, treat in a certain way or influence the status or behaviour of an individual,"* then regardless of what it is, it is personal data (Stalla-Bourdillon and Carmichael, 2018). To continue with the OpenActive example, this

could be data about exercise class timetables. Although a timetable can be published openly, once it is used to create a personalised service for an individual to suggest they attend the class, based on their preferences, it becomes personal data.

Re-identification remains a risk with open data, especially when triangulated with other data. Even when it is not possible to directly identify an individual from given information, it may subsequently become possible, based on how the data is processed, what other data it is processed with and also the means reasonably likely to be used by any person to identify a given individual.

Pseudonymisation is the process of replacing the true identifier - such as a name or ID number - with a code or other disguised identifier, and anonymisation is a process whereby the identifying information regarding the data subject is manipulated or concealed (Esayas, 2015). These can theoretically alchemize personal data into non-personal data.

The judgement of whether pseudonymisation is sufficient to do this is based on the risk of re-identification, as above. If there is no risk of re-identification, then pseudonymisation techniques may be sufficient to render the data as not personal data. Mourby et al (2018) give the following example: if pseudonymised data is held by Research Centre A, and subsequently shared with external Researcher B, if it is 'reasonably likely' that Researcher B cannot re-identify the data, these shared data are not personal data for Researcher B.

While anonymisation is theoretically sufficient to ensure data is no longer personal, Rocher, Hendrickx and de Montjoye (2019) were able to construct a re-identification model that found that "*99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes*", regardless of completeness of the data. This aside, it is likely that the very process of applying anonymisation or pseudonymisation techniques counts as processing the data (Esayas, 2015). Therefore, this process would be in conflict with the GDPR unless there was consent for the specific desired application.

Hence, it is difficult to use even anonymised data as open data. As open data requires 'all the uses all the time', it is vanishingly unlikely that such wide-ranging consents were initially required (and even less likely that another basis for processing, such as public interest, would pertain).

Sensor data complicates the issue even further. Cisco estimates that the proliferation of Internet of Things connected devices produces 5 quintillion bytes of data every day, so issues around sensor data are substantial. Advanced profiling and triangulation methods may mean that even 'innocent' sensor data (such as air quality or water monitoring data) can be compromised as personal (van Zoonen, 2016). Apparently innocuous sensors in personal consumer devices – such as gravity, ambient temperature or air pressure – can be used to infer *"highly sensitive information*" about not only the owner of the device but others in their vicinity (Kroger, 2019). Kroger (2019) further argues that such results challenge the adequacy of current sensor access

policies, and claims most data captured by smart consumer devices should be classified as highly sensitive by default, which it currently is not. The Mauritius Declaration (2014) of Data Protection and Privacy Commissioners states that, given the quality, quantity and sensitivity of Internet of Things (IoT) sensor data, identifiability is "*more likely than not*" and therefore all IoT sensor data should be treated as personal. However, this is not necessarily the case in practice.

2.5.1 Conflicts with Principles of GDPR

The six principles of GDPR are: lawfulness, fairness, and transparency; purpose limitations; data minimization; accuracy; storage limitation; integrity and confidentiality and accountability and compliance. The principles of open data may come into direct conflict with three of these: purpose limitation, data minimisation and storage limitation.

Open data, in principle, means it is impossible to implement effective purpose limitation. In the European Data Protection Supervisor's Opinion on the amendment of the directive on the reuse of PSI, it is noted that the innovation aim behind open data is explicit that purposes are not clearly defined and cannot be easily foreseen. Conversely, personal data collected for a specific purpose should not later be used for another purpose, (unless certain conditions are met). The Opinion comments, *"It is not easy to reconcile these two concerns.*"

Data minimisation is the principle that data held should be 'adequate, relevant and limited to what is necessary for purpose fulfilment'. This conflicts with 'open by default', and the aim to publish as much as possible. In the words of the Opinion, *"open data projects take accessibility to a whole new level."*

The final area of conflict is that of storage limitation. This essentially states that personal data must not be kept longer than the period for which it is required. Open data has no terms and conditions attached that specify how long it may be used for, nor would it have any way of following up on compliance with those terms and conditions. GDPR does not prescribe how long personal data may be stored for, rather, it is a risk-based approach, which requires that data processors consider the potential for risk and how to minimise it when storing data (Stalla-Bourdillon and Carmichael, 2018)

While it is important to remember that GDPR is only relevant to personal data and therefore open data publishers and users should not find purpose limitation, data minimisation or storage limitations arduous, it is clear there is the potential for grey areas around exactly what might constitute personal data and what might legitimately be re-used - especially when this theory is put into practice. It can also be seen that the GDPR has the potential for eroding the possibility for some of the most useful data to be published openly. This is not necessarily because the law has changed so significantly, but because of the scale of the potential sanctions. Welle Donker and van Loenen (2017) claim that the lack of knowledge regarding the necessary adaptations of

sensitive data to make it suitable for open data publication is a barrier to publishing high value data for many organisations.

2.6 Data Ownership

It is important to note that, the is no clear right of ownership over data. Organised datasets can be subject to intellectual property rights such as database rights, but raw data is excluded from traditional property rights (Banterle, 2018). The traditional tools of copyright, trade secrets and data protection laws are challenging to extend to raw data, and in the most recent Data Strategy the EU has made no steps to change the status of data ownership. Consequently, contractual schemes and technological access restrictions that enhance the ability to control data have dominated.

2.7 Licensing

The license is at the heart of the use of open data. Without the license, however publicly available the data may be, it is not open data. This derived from the Free, Libre and Open Source Software (FLOSS movement), when a practice known as 'software hoarding' threatened to impede the free flow of knowledge and software amongst the FLOSS community. This required a license that imposed a 'share alike' requirement on users of open source software, and the concept of the 'copyleft' license - one that facilitates sharing - was created (Stallman, 1985).

The most well-established range of such licenses is the Creative Commons licenses. For open data, Creative Commons licences CC0, CC-BY and CC-BY-SA apply. Many countries have a national or sub-national open data license, such as the UK Open Government Licence or the Flanders Open Data License. CC-BY 4.0 is interchangeable with the UK OGL, on which many licenses are based. There has been a certain amount of consolidation after an early rush of national development, yet there are still sufficient for the Open Data Portal Watch to publish a list of 'Top 10 Licenses'. Every time a license forks (a new version is created) in any way, this increases friction that not only causes problems for reusers, but also slows publication (Dodds, 2016). Van Loenen, Janssen and Welle Donker (2012) identify the variety of non-standard geodata licenses, which are challenging to comprehend for humans and machines alike, as a major barrier to the sharing of geodata.

Once data with more complicated terms is introduced into the ecosystem, it must also be accompanied by some mechanism for ensuring those terms are complied with. While there is no satisfactory automatic use tracking solution, this will require an investment on the legal (and possibly technical) side. Recovering that cost by charging for data – while compatible with the terms of open data - requires an investment in the relationship with the individual(s) or

organisation(s) who are using the data. This can be quite time consuming, and reduces the economies of scale that are possible with open data.

To conform to the Open Definition, there are only two restrictions an open license can put on reusers - attribution (CC-BY) and share-alike (any derived content or data must be published under the same license) (Dodds, 2013). In practice, licenses can proliferate while still conforming to standard definitions of open data. For instance, the Flanders Open Data License, while recommending the use of the CC0 license, also has 4 co-licenses. These include a Free Open Data License, in which the publisher retains intellectual rights of the data, and an Open Data License against a Reasonable Charge, which enables payment for the open data. The last two licenses are made to complement each other, and limit free reuse to non-commercial purposes, and states commercial reuse must be paid for. Eaves (2013) argues that in practice the variation in licenses means that open data is not a binary, but a range. Davies (2015) finds, in a study of open data for development, that emerging economies, in particular, varied in their attitudes towards the importance of attaching any license at all, and some felt that publication was sufficient indication of availability for reuse. Longshore Smith and Seward (2017) extend this, and argue that, as open licenses are based on copyleft, they make little sense in cultural contexts that originally had no copyright.

'Share-alike' licensing can have a chilling effect on the entrepreneurial, commercial use of data. Share-alike licensing has the function of promoting crowdsourcing and stimulating collective data generation and sharing. This can be illustrated with a well-known example, that of Open Street Map (OSM). OSM has always had a 'share alike' restriction, which states, "*You can basically do whatever you want with our data inside your own home or organisation, but if you then publish your results as data, then you need to tell folks about us (attribution), and share back to the public any improvements that you made, (share alike).*" This is slightly more restrictive than 'freely used for any purpose'. The implication of this for the creation of services and products is that even if a service or data is charged for, to comply with the share-alike license, third parties can then redistribute the service or data without payment.

An example of the variation that can be found in the provision of open data is offered by the edge case of Zoopla, the property data company. Zoopla run an open API, an interface that allows frequent access to machine readable data, in this case, details of properties for sale and rent. APIs require a key to use, which can be obtained by registering with Zoopla. Part of the registration involves submitting detail on what kind of application will be built with the data, and what it will do. Zoopla can turn down the application if it feels that it is too similar to other applications which already access the API, or that it will bring Zoopla into disrepute, or compete with it. Applicants must also agree to the Terms of Service, which include a time limitation on the API license of 3 months, and they must also display the 'Powered by Zoopla' logo on their site. Is this, in fact, open data in any way? Zoopla select who can access it, for what purpose, limit the license (which is for

the key, not the data) and add a requirement to sign up to the terms. They even dictate how often the data should be updated. Yet – it is also more similar to open data than any other models in existence. Anyone can ask to access the API, the licence can be perpetually renewed, and in reality, it is highly likely that free use of data would be for very similar reasons as Zoopla's constrained uses. (In reality, data sets other than geodatasets are generally used for quite specific purposes.) And the data is free, in terms of there being no cost to the user, so it seems perfectly sensible that Zoopla should manage costs at their end (eg volume of calls to the API) by ensuring as little replication of use as possible.

2.8 Data Discovery

Ubaldi (2013) notes that the creation of open government portals is the main governmental initiative in this space. These have been "*celebrated for their role in opening up and enabling the discovery and reuse of official information,*" with multiple examples catalogued by the International Open Government Dataset Search (Gray, 2017). In the UK portals are prolific. As well as transactional data on data.gov.uk, the central catalogue, there are other portals such geoportal.statistics.gov.uk, and a wide variety of local government mandated and volunteered data sets and locations published on a variety of platforms, including blogging software. Reference data is largely published outside of the catalogue, for example on informational pages elsewhere. The incentive to publish across Europe is largely regulatory, in that it is mandated by policy. It is considerably harder to find portals for non-governmental open data, although they do exist, such as the Thompson Reuters PermID project10 and Data.southampton.ac.uk (Cox, Milstead and Gutteridge, 2015). Combinations of the both, such as the Copenhagen City Data Exchange, also exist.

A key issue in both government and non-government data is that of discovery – ensuring potential users are aware of and can find the data they need (Walker, Frank and Thompson, 2015). One solution is to bring portals together, for instance the Pan European Data Portal – now data.europa.eu, itself a combination of the European Data Portal and the EU Open Data Portal (Shadbolt, 2010). Another initiative is that of creating 'data factories' in open online communities, however, as yet these offer no solution to the querying (discovery) problem. Data that is primarily useful agglomerated nationally, but which is published regionally across different authorities can therefore be delivered in a variety of formats. Hivon and Titah (2015) suggest that data needs to be available through a central central window, at least at the municipal level. The advent of Google Dataset Search, now integrated into the main search function, may change these discovery issues in the near future (Canino, 2019).

Sasse et al (2017), find that portals have low sustainability on a number of dimensions, including their governance, financing, architecture, operations and progress measurement. On the other

hand, grouping multiple datasets together in one location may facilitate the adoption of common standards, which is important to ensure interoperability. For example, using the Data Catalogue Vocabulary Application Profile for data portals in Europe (DCAT-AP), common metadata standards can be applied across multiple data portals, which can subsequently enable a cross-data portal search for datasets. This can be accomplished by the exchange of descriptions of datasets among data portals.

Based on Tim Berners-Lee's 5-stars scheme for Linked Data, the Open Data Institute's Open Data Certificates aims to assess and certify data portals that meet standards for publishing sustainable and reusable data. The ODI particularly promotes the use of open standards from W3C, the World Wide Web Consortium, to ensure interoperability among datasets.

However, a portal or catalogue is not a solution to the provision of open data, nor is it an efficient marketplace for suppliers and consumers. Ubaldi (2013) comments that "*today, many governments focus on the development of a national OGD portal as if it were a higher priority than developing technical infrastructures to open up public data for others to use*" and that as a consequence, "*the pace of change and real value will be held back*." Leonard (2012), in a critical paper, suggests that in many cases portals are no more than "*virtuous data dumps*". Governments have incentives to publish non-value adding datasets, as these are less likely to provoke critical questions or insight (Janssen, Charalabidis and Zuiderwijk, 2012). Although bodies such as the European Commission are now seeking to understand which data sets comprise those of 'high value' there has not necessarily been a relationship between data that is important to a region or country, and data that is published. In previous work, I have shown with a statistical assessment of the open data agriculture datasets harvested by the European Data Portal that there is no relationship between the importance of agriculture to the GDP of a country and the number of agriculture datasets published by that country (Walker, Thuermer and Simperl, 2019).

Davies (2010) and McLean (2011) argue that the view that government's role is to remove technical (via publication online) and legal barriers (via appropriate licensing) is too simplistic, and that to achieve the aims of open data publishers may need to take a broader view of activities. These may include the creation of new datasets, the devotion of extensive political and social resources and the support of interventions. In some situations, government may attempt to make the data available in app form, such as those available from the UK's Ordnance Survey. However, it is far from clear that this meets demand better than the raw data provision, and a Code4Kenya (a government outreach initiative) app development project as a stimulant for demand resulted in very poor levels of use (Van Schalkwyk et al, 2015).

While they are relatively easy to implement, there is no evidence that these top–down approaches address users' most pressing concerns. As such, they are weakly linked to the impact of open data. Research is beginning on user-oriented provision of data. The EU has commissioned

research into 'High Value Datasets' – data that can be used across multiple domains. Walker, Taylor and Carr (2015), building on the work of the Open Data User Group (2015) examine how providing open government data in a 'joined up' format rather than as separate datasets will enable greater use in key areas such as health, transport and construction. It suggests that co-locating data provision with tools and services, rather than a simple platform, catalogue or portal, will reduce the data capability requirements demanded of the user and support increased usage. Walker and Simperl (2018) establish a 10 point, user-oriented roadmap for increasing usage, including borrowing strategies such as recommendations and reviews from e-commerce.

The London City Data Strategy (2016) recognises this challenge, and states that one of its core principles is that, "*non-technological elements and technical domains will have equal status. Business models, value networks, feedback loops, a data marketplace, data toolsets, clear licensing arrangements, and more efficient data governance, are all vital building blocks of a functioning city data economy*."

As Hivon and Titah (2015) note, "*data becomes valuable when it is used, not when it is published*" and in the next section usage is explored in more depth.

2.9 Data Reuse

Having discussed the supply side of the open data equation, the demand side will now be addressed. There are a great number of potential benefits of open data when it is used. An overview of a number of benefits of open data identified by interest groups and in papers are shown below.

Table 2 Benefits of Open Data

Benefit/ Source	Democracy	Policy	Efficiency	Transparency	Society	Economic	Research
Opengov data.org	Increased civil discourse	Improved public welfare	More efficient use of public resources				
Shadbolt (2010)		Evidence based policy	Accountability	Transparency	Social value	Economic Value	
Sunlight Foundation			Accountability	Transparency			
OKF	Participation and Engagement			Transparency	Social value	Releasing commercial value	
Data.gov	Increased public participation in	Informed policy	Cost savings, efficiency, improved civic services, accountability	Transparency		Fuel for Business	Research and Scientific Discovery

Benefit/ Source	Democracy	Policy	Efficiency	Transparency	Society	Economic	Research
	democratic dialogue						
Open Data User Group (2015)	Citizen interaction with services	Evidence based policy making	Improved design of services, Smart Cities, IoT	Transparency		Growth and Innovation	
OECD (Ubaldi 2013)	Promoting citizen engagement and social participation	Creating empowered civil servants	Fostering efficiency and innovation in public services, accountability	Transparency		Creating value for the wider economy	

With such a wide variety of benefits, as Davies (2010) writes, open data holders are, "*unlocking potential while being essentially agnostic about the sorts of potential unlocked."* However, there are key foci of value. Chui, Farrell and Jackson (2014) identify three value levers; decision making, new offerings and accountability. New offerings particularly refer to non-governmental bodies and individuals; "*When the government and other stake-holders release data, they help companies, agencies, and individuals to develop innovative apps, products, and service.*" Chui, Farrell and Jackson (2014). Susha, Johannesson and Juell-Skielse (2016) note that although there is now a developing corpus around the critical success factors for publishing data there is no such holistic framework in the research for data usage. Tennison (2012) states; "*I find business cases for data publishers much more compelling than examples of how open data can be used."* Ferro and Osella (2013) claim that PSI re-use by private sector entrepreneurs is "*struggling to take off due to numerous inherent roadblocks"* and, *"vagueness surrounding the rationale of underlying business endeavours".*

McLean (2011) cites the original emphasis on the reduction of institutional barriers to access to open data as encouraging an "*overly-optimistic*" view of how widespread use would be by individual citizens. The UK government, when opening datasets, acknowledges that open data should '*stimulate innovation'* and that '*demand for data and broader market forces*' should be the main driver. Despite this, Dulong de Rosnay and Janssen (2014) argue that government can benefit directly from the use of its own data, whether that be from improving decision making, creating better public services or simply improving their internal data management practice.

Whether the right data is released - as measured by outcome or impact metrics - is important, and this is still underexplored (Brandusescu and Lämmerhirt (2019). Reduction in Freedom of Information requests was an early assessment metric (e.g. Lee and Kwak 2011) but this does not address the key issue of new businesses created with open data. Harrison, Pardo and Cook (2012) found that government officials had *"limited understanding of the economic implications of adopting open data as a new line of business.*"

Lassinantti (2014) argues that any open data service, whether directly regarding the public sector or not, creates value for government *"When the developers create value for themselves in the form of new revenue streams (companies), free services for their local community (developers) or educational web services (journalists), values on a higher societal level are also created for municipalities; regional growth, new business opportunities for the citizens and empowered citizens are simultaneously created."*

2.10 Use and Impact

It is clearly important that it is understood *"under what circumstances they [data consumers] will be best equipped to make use of it [open data]"* (Harrison, Pardo and Cook, 2012). Erickson et al (2014) posit that users may encounter problems understanding what the data in a set comprises. Davies and Frank (2013) attempted to use a data set and were challenged by, *"our consequent discovery of the layers of understanding we needed to comprehend how the data could be re-used.*" At the core of this challenge is the fact that the data collected is a by-product of another process entirely. This disconnect suggests it is difficult for a reuser to ascertain the uses to which data may subsequently be put without reference to the publisher. Davies (2012) proposes a non-technical complement to the '5 stars of Linked Open Data', which he designates the '5 stars of Open Data Engagement'. In the third to fifth stars the engagement becomes dialogue between user and publisher. The user is increasingly important in open data as the emphasis shifts from supply driven provision to demand led opening (Van Loenen, 2018).

Projects such as the Open Data Impact Repository acknowledge that little is known about use and impact and seek to gather examples and case studies to create an evidence base. A major challenge is how to quantitatively assess usage. Lee and Kwak (2011) note that agencies rely on process-centric metrics to assess engagement, such as how many people have downloaded, or requested, data. While the number of applications based on open data, or the number of open APIs, may be an example of demand in some sectors, such as in London where the government department TfL no longer makes its own apps, but largely depends on the private market to fulfill this role, in other areas it can be interpreted more as government level promotion (a legitimating process) than actual proof of use (Hivon and Titah, 2015).

There are many motivations for measuring open data. It is needed to maintain quality of data and support; to justify investment; to focus resources to most effect; to compare progress between countries, institutions and portals and to set benchmarks for countries, institutions and portals. However, it is still difficult to know exactly what to measure and how to measure it. Once a metric is decided, it becomes the focus of both effort and observation. This can detract from other important aspects of assessment, and can also result in 'target chasing' - investing time and resources to affect a specific suite of metrics to the exclusion of others (Frank and Walker, 2015).

Even before any technical issues of measurement can be addressed, there are broader issues that affect any attempt to define a set of metrics, even in areas that appear to draw consensus. For instance, 'quality' is often used as a metric for data. There is a substantial body of work applying key elements of 'data quality' to open datasets and assessing them in that manner (e.g. Vetro, 2016). However, this term can mean very different things for different types of data and different types of users. Is a good quality data set one that has all fields completed, or that perhaps has fewer fields but is more accurate? Is a good quality data one that has been cleaned, or one still contains the raw data, including outliers (Gomez-Cruz and Thornham, 2016).

Secondly, there is often misalignment when deciding who 'the users' are that any metric should attempt to measure. Is it the primary, secondary or tertiary users that need to be assessed? What particular aspect of reuse activity should or could be measured - downloading, integrating, creation of application or use - and how can this be addressed when different users perform different functions? Further there is not clear track between a portal and use - even if data appears in an app, it may have actually been collected from one of many sources - the original data, a copy, or via a catalogue. Boswarva (2015) notes that, as government open data strategy moves towards the API, facilities such as bulk downloads, which may be more appropriate for some groups of users, are abandoned. It is not incredible to consider that commercial data, more influenced by what is possible to achieve technically and with less incentive to maintain more low-tech access, will experience this to an even greater degree. The nature of portals that might require data to be patched together from several sources means that it is difficult to see what the real demand for a data set is (Davies, 2010). Data sets of this nature may not show much use because of infrastructural challenges rather than because of their content.

The characteristics of open data that make it difficult to measure are well-documented; it is non-rivalrous, has strong network effects, is related to a philosophy that rejects tracking, and cannot be tracked through financial payments. It is challenging to isolate and disambiguate 'use' and 'impact', further, the range of possible subjects to measure is vast. This leads to confusion over what needs to be measured. Impact, especially, can be as diverse as participation rate at elections, economic growth, improvement of reliability/efficacy/speed of public services, reduction in the number of homeless or increase in engagement in participatory budgeting (Walker et al, 2020). The most common methods for measuring use and impact are user surveys; business population studies; microeconomic studies and macroeconomic studies.

How businesses and citizens might use – or be prevented from using – open data will now be discussed in more detail. The highly technical nature of open data practice harbours the potential for citizen users to become disengaged from the process of shaping and constructing relevant quality characteristics.

2.11 Open Data Users

Hellberg and Hedstrom (2015) suggest that the "myth" of open data is that while people like the idea of open data, this does not translate into active participation in the reuse process. Indeed, other authors have critiqued the lack of co-located tools, processes and skills guidance with government data repositories (Walker, Taylor and Carr, 2015; Frank and Walker, 2015). DiCindio (2012) argues for the necessity of 'deliberative digital habitats', online spaces for engagement.

To accurately identify who wishes to use, or is using, open data and how they access it, the people and businesses (agents) and processes involved must be explored.

There have been multiple attempts to define useful groups of open data interactants, based around open data value chains, ecosystems and business models.

2.11.1 Open Data Value Chains

The value chain is traditionally the process by which an organisation adds value to a product or service. In exploring the 'open data value chain' authors such as Ferro and Osella (2013) have focused on activities carried out with open data and assigned organisations and individuals to each stage.

Table 3 Open Data Value Chain (Ferro and Osella, 2013)

PSI Generation and Dissemination	PSI Retrieval, Storage, Categorization and Exposure	PSI Re-use	PSI Consumption
Government Bodies	Non-profit enablers	Core reusers	Business end-users
	Private sector enablers	Service advertisers	Government end-users
		Advertising factories	Consumers
		Spontaneous civil initiatives	

More recent work has moved away from interpreting open data through the traditional value chain lens, perhaps because the concept of a chain implies a vertical control and a hand-off from one activity to another that is not appropriate to open data. Hivon and Titah (2015) focus simply on the activities performed by citizens in engaging with open data.

Table 4 Citizen Engagement with Open Data
(derived from Hivon and Titah, 2015)

Activity	Description
Identifying	
Requesting	Either in person or via systems
Converting	Data into raw format
Programming	Data downloading to publication of apps.
Promoting	Via completed apps

Hivon and Titah's (2015) structure invites many questions around the first two activities, which are often opaque to observers and research. They also refer to citizens who 'identify' open data

as ‘champions’ and this immediately suggests that even the theoretically simple task of identifying useful open data requires some specific skills. Frank and Walker (2016) support this, finding that the workshop participants they engaged with regarded data identification as onerous and challenging. The possibility of requesting via a person provokes the question of how it is possible to make access via a person rather than a system equal for all citizens. Unlike the other four activities, whose satisficing are largely in the hands of the citizens, requesting implies a satisfaction of the request from the data holder. This activity, therefore, can be seen as problematic in the citizen value chain. Indeed, Hivon and Titah (2015) note that, “*Citizens do not feel they have control over the released data*.”

Interestingly, civic advocacy groups in this study felt that they had more influence but there is no evidence to inform whether the types of data were the same in both cases. Ferro and Osella (2013) differentiate between ‘profit’ and ‘non-profit oriented actors’ in their value chain. However, it may be argued that this is an artificial distinction. A social enterprise may have more in common with a commercial business than a civic activist.

The London City Data Strategy attempts to extend the possible interactions with data while being less prescriptive about whether business, government, commercial interest, third sector or individuals are appropriate to these roles, having only three categories: data enrichers, integrators and consumers.

2.11.2 Open Data Ecosystems

The ecosystem is an attractive model for theoreticians and practitioners for several reasons. Firstly, the ecosystem metaphor has successfully described the interactions and cycles and flows of resources and information that one needs in describing an organisation, to the point that ‘ecosystem’ is equally at home with the prefixes ‘business’ or ‘organisational’ as suggesting an ecological setting. Although it has been suggested that this metaphor has flaws (Mars, Bronstein and Lusch, 2012), ecosystem metaphors are not barriers to the use of other theories.

Heimstadt (2014) creates an “Open Data Ecosystem Taxonomy” where data flows from Suppliers through Intermediaries to Data Consumers and then feeds back to Suppliers. This loop is supported by Enablers; Policy, Infrastructure, Capacity, Funding and Knowledge. Sanderson (2013) offers up entities (data, information) and agents (government, infomediaries and others). He does, however, suggest that the system is more dynamic and interrelated: the government does not simply push data to infomediaries who in turn push it to consumers, but more direct participation can take place. Neither of these models explores the possibility that the most significant two-way interaction might be conceptualised as existing between the inter/infomediaries and the data suppliers.

Davies (2010), despite pre-dating the work of Heimstadt (2014) or Sanderson (2013) creates a much more detailed model of an ecosystem, noting the interrelationship of data journalism, integration in third party databases, mobile apps, APIs and more. He argues that community building and facilitating skills are vital parts of an ecosystem. These may be filled entirely by third parties, with government as overseers.

2.11.3 Open Data Business Models

The term 'business model' has no single agreed definition, but can be described as the set of (operationalised) assumptions an organisation holds about its key activities, revenue streams, customers, costs and value proposition (Ovans, 2015).

Many authors base business models for open data on those of open source software (e.g. Tennison, 2012). Both Ferro and Osella (2013) and Zeleti et al (2014), whose work builds on that of both Ferro and Osella and Tennison, cite 'open source' as only one among many potential business models.

Ferro and Osella (2013) differentiate partially between business models because of their position in the value creation process. How real, however, given the problems of applying the value chain to open data, is this distinction? Tennison (2012) mainly divides between revenue streams and cost reduction, although both of these may exist together in the same model. Welle Donker and van Loenen (2016) identify no less than 17 separate possible business models for sustainable publication of public sector open data, although they acknowledge that not all are suitable. However, it is rare to find examples of all but one of these models operating in practice. All the national portals of the European Member States operate a budget financing model (Open Data Maturity, 2017). There are occasional examples of successful freemium models, such as that established by the Finnish Meteorological Institute (Walker, Hewitt and Simperl, 2020).

Zimmerman and Pucihar (2015) use the case studies of three open data businesses and apply the Business Model Canvas (Osterwelder, 2004) to highlight their activities. However, the Business Model Canvas was designed for the entrepreneur to complete, and it can be argued that it is difficult or impossible to know or infer all of it accurately from outside the company. Nor does it provide a model for further investigation. Zimmerman and Pucihar (2015) go on to assess their three case study companies against Ferro and Osella's 'archetypes', but suggest one company fits both the 'open source like' and 'freemium' categories; either, therefore, the archetypes are not sufficiently discrete or the analysis of the company is insufficiently focused.

The most empirical of these papers is that of Lindman, Kinnari and Rossi (2014), who interviewed 14 active Finnish app development companies. Their identification of these companies based on the dimensions of Rajala's (2009) framework is shown below. As apps are frequently end-user oriented, it may be that their sample was biased towards this specific 'value network profile', as

they refer to it, however, the assignment of the attributes is still a valuable contribution in a largely theoretical area.

Table 5 The Role of Intermediaries
(derived from Lindman, Kinnari and Rossi, 2014)

Value network profile (#)	Offering	Revenue model	Resources	Relationships
Extract & transform (1)	Find and convert raw open data into a format allowing further analysis and processing.	No revenue model, the organization operates pro-bono	Open data and the volunteer developers working on the toolkit	Open source community, open data community, and public administration and private data publishers
Data analyzer (3)	Create visualizations or algorithm-based analysis to generate new knowledge from the data.	Project work, product based transaction pricing, and modular ecosystem	open data, extracted & transformed data, and private or commercial data	Open data community, public administration, and global information providers
User experience provider (7)	Create interactive user interfaces with help of open data sources.	Advertisements, one-time fee, subscription, donations, licensing and freemium	Open data, extracted & transformed data, scraped data, private or commercial data	Open data community and subcontractors
Commercial open data publisher (1)	Publish data to be freely used by the community	Cost saving through crowd-sourcing of new user interfaces	Maintenance and updating of the API towards the published data	Open data developer community to increase awareness and usage of the data
Support service and consultation (2)	Supports other companies in the open data value network.	Project work and service-based pricing	In-house programming, consulting and subcontracting	Open data community and small open data companies as innovation partners for large technology companies

Suggestions that big and open data have the potential to empower people are problematic (Gurstein, 2011) in that it is often those who are already empowered through having the required skills or access to infrastructure who benefit and who define the products resulting from using big and open data as material. This divide emerges the idea of the need for greater data literacy skills in the wider population to reduce this dependence, which are addressed in the next section.

2.12 Data Literacy and Skills

Use will largely be limited to intermediaries if enough interactants and potential users do not have the skills to engage. Whilst many open data initiatives are led with equitable intentions, "given the scarcity of formal data literacy curricula and initiatives for average citizens, the extent of that inclusivity is arguably minimal." (Argast and Zvayagintseva, 2016). Frank et al (2016) argue that *"data literacy (the ability of non-specialists to make use of data) is rapidly becoming an essential life skill comparable to other types of literacy"*. It is not possible for individuals (or many organisations) to rely on IT support teams to help create reports to interpret data, and it is even less likely that community groups will learn technical schema such as RDF and place data into triple stores (Dix, 2014). A 'pre-use' type of event in Toronto, the Ontario DataJam, found it

necessary not only to equip citizens - in this case, those involved in not for profits - with the knowledge of what resources they might be able to access and where to access them, but also to introduce them to the wider civic technology landscape, including presenting examples of successful Datathons. Data literacy, therefore, can be understood as not only about data skills but about the "*data culture*" (Kayser-Bril, 2016).

Citizens most often come into contact with open government data, therefore the choice of data and the areas for development and consumption of data is often directed by government. This is often led by large metropolitan areas (Argast and Zvayagintseva, 2016). A problem arising with this government-led activity is that citizens who are not engaged with the civil service or local government in some way may have very limited access to data availability. This means that data activities are occurring in a top down, rather than bottom up, or integrated, fashion. While few people have data knowledge, far more have problems that they are happy to discover may be improved with data. The Toronto node of the ODI worked with the Toronto Public Library to offer a series of workshops on data based in the library, and also asked how the library could be used to make communities more resilient, connected and successful (Argast and Zvintseyeva, 2016). This approach emphasizes the social as well as the technical - a 'low tech' intermediary is established as a new route to data. However, these intermediaries are increasingly unavailable in less densely populated areas.

2.13 Summary

Despite being a relatively new field of research there is a vast amount of open data scholarship covering a wide range of subjects. Regardless of overarching directives and specific mandates, deciding exactly what to open and how is still a challenge for most data holders. This challenge is partly because of resource constraints, but also because of potential technical, economic, legal and political risks. A major issue is privacy risk. While personal data can never be open data, as more forms of data, especially those created by the Internet of Things, are created, and as more types of application are developed, understanding what might be personal data has become increasingly complicated.

Once data is successfully opened, there is a further challenge – how to ensure it is used. Many studies have pointed to open data's potential to support substantial economic growth, so the interaction of business and developers with open data has been well researched, resulting in clear models for open data value chains and business models. Although citizen engagement is posited as a legitimate use for open data – especially open government data – often a lack of skills or lack of awareness of civic technology means this not such a fertile ground for use.

Increased use for innovation purposes would increase incentives to open data, and could also, by boosting tax revenues, providing employment and improving services, potentially reduce resource constraint on publication. Measuring innovation use is difficult, because of the nature of open data as available for anyone. There are ways to make it trackable but these conflict with the spirit of open data, which is as much a part of the 'openness movement' as a way to ensure data is available. However, in some cases, open data initiatives, such as that of Zoopla, push the boundaries of what is considered open data in order to achieve sustainability.

Next I explore the mechanisms by which economic value can be created with open data through innovation.

Chapter 3 Literature Review: Open Innovation

As discussed in the previous chapter, open data is seen as a valuable tool of innovation. In a quantitative analysis of measurements from 61 countries, Jetzek, Avital and Bjorn-Andersen (2013) hypothesize that openness, "*positively affects the ability of society to generate value from data through the innovation mechanisms"* and that, "*data-driven innovation positively affects value through generation of new knowledge, new processes, services and products, and new businesses".* They find they can support both hypotheses, with the second hypothesis being highly significant and having a high absolute value.

3.1 Innovation

In the early part of the twentieth century, Austrian economist Joseph Schumpeter popularised the idea of innovation as a driver of economic growth, stating that anyone seeking profits must innovate. He identified five types of innovation: new products, new processes, new markets, new supplies of materials and and new industry structure (for instance, moving from a monopoly to competition (Śledzik, 2013). Schumpeter also placed great emphasis on the dual activities that combined to create innovation: an idea must be both discovered (invented) and executed upon (by a dynamic entrepreneur) in order to constitute innovation (Schumpeter, 1934, in Sledzik, 2013).

Innovation studies in practice have been dogged by questions around whether the products, processes, markets, suppliers and structures in question are sufficiently 'new' to be 'innovative'. Measuring innovation is not simple, and a test for 'innovativeness' is hard to establish. A test for innovation in an open regime is even more challenging, as there is little or no way to track it. This research does not seek to assess whether 'innovativeness' of a quantitatively new product or service has been achieved, but investigates innovation as a process, using Sørensen and Torfing's (2011) definition of, "*an intentional and proactive process that involves the generation and practical adoption and spread of new and creative ideas, which aim to produce a qualitative change in a specific context*".

Innovation theories, especially popular ones, further have the propensity to suffer from 'scope creep'. This has two drivers; the first is that the finer points of the theory may be poorly understood in the wider community through which it has disseminated, and the second is that if the theory is sufficiently exciting, practitioners and researchers alike are keen to demonstrate that they have effected or identified examples of it. An example of this is disruptive innovation, where many more new products or processes are claimed to be disruptive than fit with the original model (Bower and Christensen, 1995;, Christensen, Raynor and McDonald, 2015).

The traditional model of innovation is one where a business (or more rarely, an individual) generates, executes, markets and distributes new products or services. The expense of the investment into generating and developing ideas - only some of which will make it to market - is recouped by ensuring the ensuing financial benefits are returned only to the innovating company, and not competitors. This requires close ownership of the idea and relies on the concept of intellectual property to effect this, whether this takes the form of trade secrets, patents, industrial designs or copyright (Gutzmer, 2016). As a result, large firms with extended research and development capabilities and complementary assets could outperform smaller rivals (Teece, 1986).

In the context of open data, therefore, the traditional model is not a good fit. As previously noted, a substantial share of open data is published by national, local or municipal government. Such publishers have historically not been well-positioned to discover new value in their own data. This may pertain to cost, competing calls on time, a lack of insight into where value might lie, or often because the processes and skills required for the production and collection of data are not those required to exploit the data.

Further, and crucially, it has been argued that the exploitation of open data is not an appropriate function of government, as this deters competition and weakens innovation (Yiu, 2012; Gurin, 2014) and that engaging the private sector in re-usability and innovation is critical (McLeod and McNaughton, 2016). An appropriate model for investigation, therefore, is open innovation.

3.2 Open Innovation

Open innovation is a model of industrial innovation first identified by Chesbrough in 2003, and over the past decade and a half it has become a key concept in innovation studies (Chesbrough and Bogers, 2013). It is defined as *"the use of purposive inflows and outflows of knowledge to accelerate internal innovation and expand the markets for external use of innovation, respectively"* (Chesbrough, 2011). In 2013, Chesbrough and Bogers extended this to, "*a distributed innovation process based on purposively managed knowledge flows (spillovers) across organizational boundaries, using pecuniary and non-pecuniary mechanisms in line with each organization's business model*". This reflected concerns that the definition did not sufficiently emphasise value capture and downstream activities, and also to establish the definition more clearly in previous economic literature.

Open innovation has its roots in the idea that 'erosion factors' such as increased worker mobility, more capable universities and increased availability of start up capital from funders combined to decrease the ability of organisations to develop innovation that led to competitive advantage through traditional innovation (Chesbrough, 2003). It also developed as a legitimate and plausible

response to the economic concept of spillovers. Spillovers occur because of the exploratory nature of a firm's research and development activities. As the outcomes of such investment naturally cannot be specified in advance, it may produce outcomes that 'spill over' beyond the ability of the original firm to benefit from them, for a variety of reasons. Technological or research and development spillovers are most often defined as externalities (Dumont and Mueesen, 2000). However, open innovation suggests that purposively structuring mechanisms can be utilised to manage such spillovers as part of the innovation process.

The central concept of open innovation is the movement of ideas and knowledge across the organisational boundaries, as shown below.

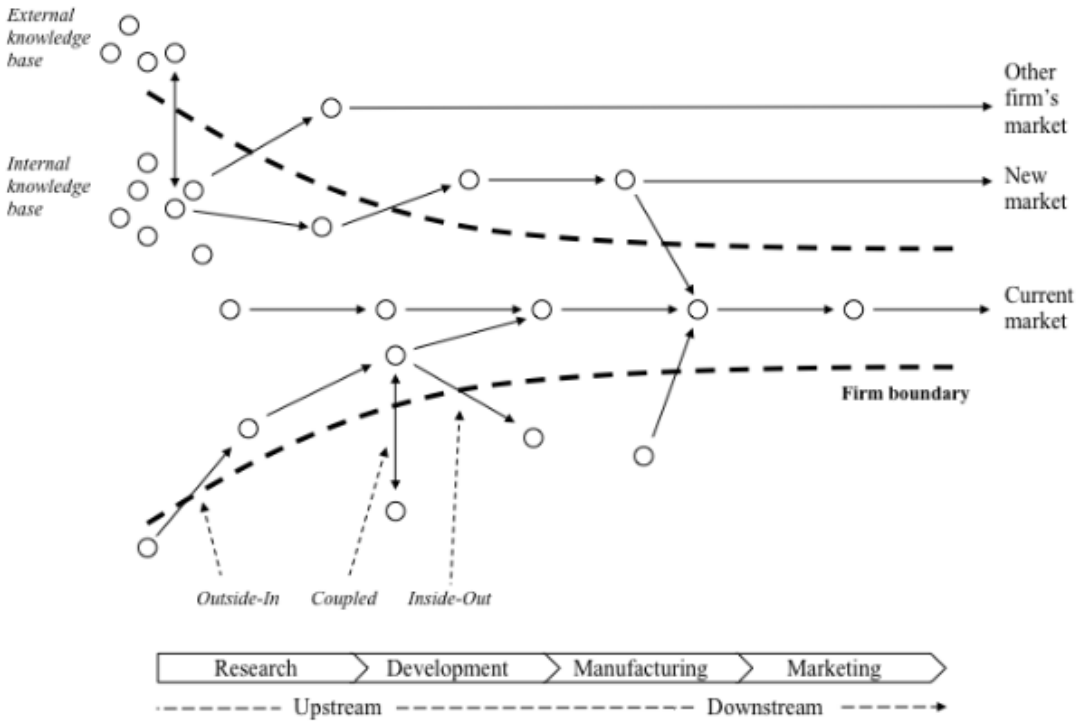

Figure 2 The Open Innovation Funnel Model (Chesbrough and Bogers, 2013)

It is a distributed, decentralised and participatory approach to innovation, which focuses on transferring knowledge from where it is created to where it can best be commercialised (Debakaere et al, 2014). For businesses, open innovation both reduces costs and allows access to disparate information that it would be impossible to capture inside the firm. This 'outside in' focus of open innovation is the most commonly recognised aspect. However, it is equally important that under-utilised ideas within the firm are distributed into the wider environment to be incorporated into the innovation processes of firms more able to exploit them. Lichtentaler and Ernst (2009) express these two sides as *'technology acquisition*' and '*technology exploitation*'. Therefore, establishing a coherent business model that defines what is acquired and what is exploited is key to open innovation (Chesbrough, 2011). There are three core processes: outside in, inside out and coupled innovation (Gassman, Enkel and Chesbrough, 2010). Outside in enriches the

organisation's own knowledge base through integrating suppliers, customers and external knowledge sources. Inside out processes occur where a firm is earning revenue by marketing products and services, selling intellectual property and transferring ideas to the external environment, often through becoming a supplier or customer of a new initiative. The coupled process consists of co-creation with complementary partners through alliances, cooperation, and joint venturing.

The majority of research in open innovation focuses on the inbound (outside in) aspects, investigating how firms leverage knowledge and technology for internal innovation. (Chesbrough and Bogers, 2013; West and Bogers, 2013). Outbound open innovation is much less researched, although there is a growing body of work on selective revealing, intellectual property, and licensing. There is still a need for greater understanding of how coupled open innovation processes work (West and Bogers, 2013). The focus of the research corpus is summarised below.

Table 6 Summary of Modes of Open Innovation

Inbound	Outbound	Coupled
Most researched	Mostly focuses on IP/Licensing	Little researched
Reduces cost, allows access to disparate information (Chesbrough, 2006)	Under utilised ideas are distributed into the wider environment	
'Technology Acquisition' (Lichtentaler and Ernst, 2009)	'Technology Exploitation'	
Enriches organisations' own knowledge base (Gassman, Enkel and Chesbrough, 2010)	Firm earns revenues by transferring ideas to external environment	Co-creation through alliances, co-operation and joint ventures

3.2.1 Instruments for Open Innovation

The two instruments by which organisations effect open innovation are the establishment of an architecture for sharing and co-creation, along with the construction of a community - the ecosystem - which participates in open innovation (Chesbrough and Appleyard, 2007). Figure 3 below shows some of the forms these might take.

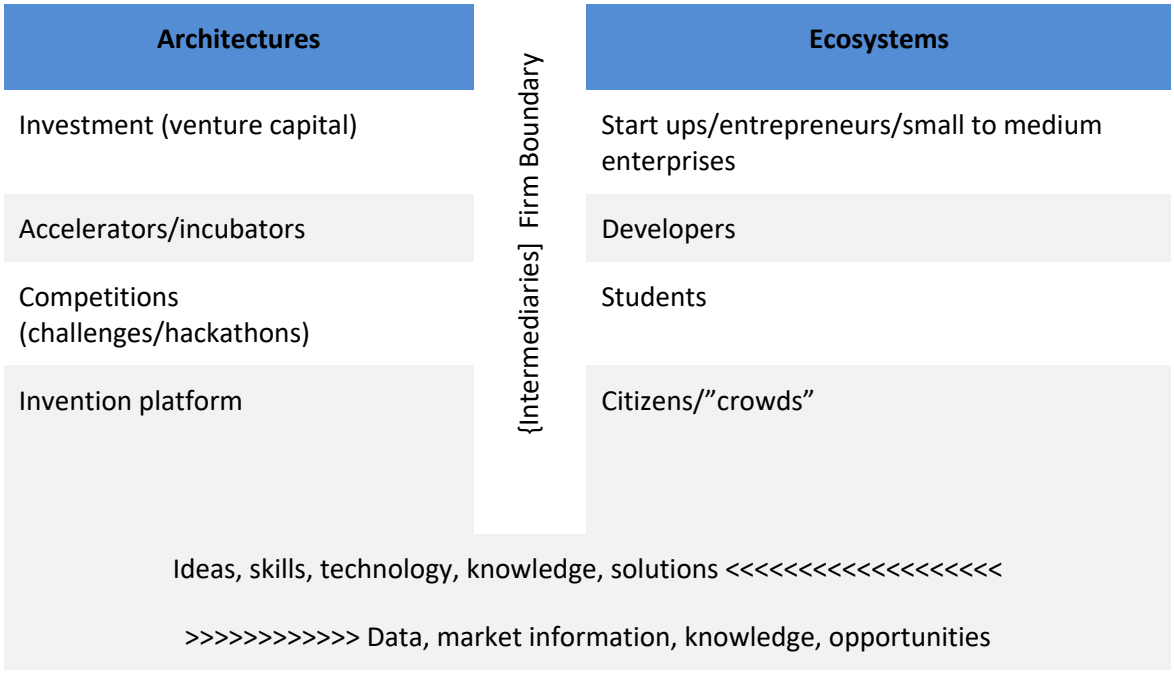

Figure 3 Common Types of Architecture and Ecosystems in Open Innovation

The ecosystem and architecture are not independent of each other; the two are very much designed in tandem in order that the architecture appeals to the ecosystem, and conversely, the ecosystem matches the architecture. Hence, an organisation that wishes to invest in start ups will expand its entrepreneurial ecosystem, and an organisation that aims to attract developers may establish a series of competitions. The Danish toy manufacturer LEGO is primarily interested in ideas that their customer base is attracted to, and run an innovation platform, LEGO Ideas, that allows LEGO fans to suggest, and vote on, new ideas.

Organisations may wish to diversify their approach. For instance, Deutsche Bahn (DB), the German state transport operator) has established DB Mindbox, an 'innovation lab' in Berlin, which is a physical and central place *"to embrace fresh ideas and creative solutions"* according to the Deutsche Bahn website, dbmindbox.com. Mindbox has three programmes: StartUpXpress, an accelerator which provides funding, coaching, mentoring and the possibility of DB as a future customer to start ups in the transport technology space; DB Challenge, a series of contests aimed at individuals or small groups from technical, software development or design backgrounds; and DB Open Data, a series of events centered around hackathons, aimed at software developers. The rewards for participating in DB Challenge and DB Open Data include networking, the potential for employment, or demonstrating or honing skills in order to take advantage of future opportunities.

3.2.2 Intermediaries, Markets and Crowds

Although some researchers conflate the ideas of open innovation as a kind of crowdsourcing (Mergel, 2017), crowdsourcing is a specific kind of broadcast search mechanism used in open

innovation (Chesbrough and Bogers, 2013; Seltzer and Mahmoudi, 2012; Miadenow, Bauers and Strauss, 2014).

Crowdsourcing is a key open innovation activity in the public sector (Kankanhalli, Zuiderwijk and Kumar Tayi, 2017). This is both because citizens constitute a major resource, and because they are frequently the stakeholders who are, or will be, most affected. Boudreau and Lakhani (2009) suggest that there are two organizing approaches for bringing in outside innovation: collaborative communities and competitive markets. Crowdsourcing is often associated with collaborative acts of peer production, such as OpenStreetMap or Wikipedia. Firms aiming to use the crowd as a source of innovation frequently use the method of disclosing innovation-related problems via online platforms, inviting external experts or users to contribute to solving predefined innovation challenges (Frey, Luthje and Haag, 2011; Jeppesen and Lakhani, 2010). Developing such crowdsourcing contests internally is certainly possible for larger companies such as Microsoft (who run the Imagine Cup) or Netflix (2009's Netflix Prize). This is more challenging for the smaller companies identified by Van de Vrande et al (2009). However, such contests are often managed by intermediaries. As the concept of open innovation has grown in importance, innovation intermediaries drew more attention (Chesbrough, Vanhaverbeke, and West, 2006; Lakhani et al. 2007). InnoCentive is one such intermediary, whose business model focuses on broadcasting science problems for 'Seekers'. It has a user base of 400,000 'Solvers'. Solving the problem is often substantially financially rewarded. Kaggle.com offers organisations an opportunity to access the data science crowd to develop machine learning based solutions.

Boudreau and Lakhani (2009) emphasize that the dynamics of community and markets are very different. When reaching out to competitive markets, the profit motive is important, relationships are formally governed by contracts and there is little sharing between external organisations. In collaborative communities, the relationships are less formal, there is more sharing of technology and intrinsic motives (developing a new skill, working on something aligned with personal values) are key. Almirall, Lee and Majchrzak (2014), in an empirical review of 6 cities engaging in civic open innovation, argue that in fact, collaborative and competitive approaches work together. They write that by creating value for external agents rather than, "*acting as passive providers of data and resources*" cities can introduce competitiveness, and companies can be more collaborative when they proactively engage with cities (Almirall, Lee and Majchrzak, 2014).

3.2.3 Open Innovation in Practice

Industry and type of activity affects the likelihood of a company engaging in open innovation. Within high tech industries such as those comprising information technology and electrical and electronic engineering, the number of joint research and development projects comprises almost 50% of all research and development projects within a company (West and Bogers, 2013). This is also affected by size of company; Van de Vrande et al's 2009 study of 605 small and medium

enterprises (SMEs) found that the larger the company, the more likely it was to engage in open innovation, regardless of type of industry. Small enterprises (up to 99 employees) were less likely to engage in open innovation. However, the authors also suggest that smaller companies can take advantage of the open innovation model by becoming part of the participating community, utilizing exploited technology or contributing to the inflows of innovation to a larger company which can more readily develop and distribute the innovation.

Laursen and Salter (2006), in an analysis of the innovation search practices of 1700 UK businesses, find that although broader search was associated with greater innovativeness, this was a curvilinear relationship, and too much searching led to diminished returns. This makes intuitive sense, as a risk of open innovation is that it overwhelms the organisation with potential innovations, which it then lacks the resources to develop meaningfully. These resources might be absorptive capacity, time or attention (Koput, 1997).

3.2.4 Critiques of Open Innovation

"*There are a lot of people who claim to talk about "open innovation", but are actually talking about something else,*" (Christensen, 2012, quoted in Chesbrough and Bogers, 2013). As with Christensen's own concept of disruptive innovation, the appeal of the theory has led to open innovation being claimed by a broad church. This has further caused dilution by imprecise theoretical definition. In an analysis of the literature Dahlander and Gann (2010) found there were many different constructions put upon 'open innovation' and it suffered from some flexibility of interpretation, evidenced by a diverse and fragmented researcher base. The edges have blurred with user innovation, open collaborative innovation, crowdsourcing (as seen above) and various other activities, as acknowledged by Chesbrough and Bogers (2013).

Other critiques focus on the concept, the model and application of the theory. Entitling their paper, 'Old Wine in New Bottles', Trott and Hartman (2009) argue that open innovation is merely a repackaging of open communications between firms. They further critique the model for being linear without feedback mechanisms, and argue that it requires a cyclical element.

Critics including Mowery (2009) and Tidd (2014) note that context is important for open innovation, implying open innovation will work differently in varying institutional environments, and may not work at all, or poorly, in some. Trott and Hartman (2009) suggest that interfirm boundaries tighten when external boundaries are loosened, and that the implications are of this are not accounted for. Further, Dahlander and Gann (2010) find a lack of research into the costs of openness, as opposed to the benefits.

Tidd and Bessant (2013) argue there are issues of vagueness around the application of theory; including how to actually search for ideas, the requirement for substantial research and development capabilities to exploit ideas once gained and potential conflicts of strategic interest.

3.2.5 Barriers to Open Innovation

Engaging with open innovation requires not insignificant resources. An implication of open innovation is that as internal innovation becomes less important the management process becomes conversely more vital. Barriers to open innovation include 'not invented here' syndrome (Chesbrough and Crowther, 2006) and various cultural, organisational, cognitive and institutional distances between collaborating organisations. Free-riding, limited resources and legal issues also present barriers.

3.2.6 Open Innovation and the Public Sector

While the larger part of research in open innovation has focused on private firms, there is also an emergent body of work investigating open innovation in governmental contexts. As well as moving from manufacturing into services, open innovation has moved into the social and public sectors, where it is known as 'open societal innovation'; "*the adaptation and subsequent sustainable use of appropriate open innovation approaches from business, adapted and utilized by state and society to solve societal challenges,*" (von Lucke et al, 2012, in von Lucke, 2014). Countries which instituted open innovation policies facilitated a positive innovation environment (Lee, Hwang and Choi, 2012).

There are key differences between the drivers of innovation in private and public sector organisations. Public sector employees rarely focus on attracting or retaining customers; few are hired to search for new markets or ideas (Mergel, 2017). It is highly plausible that the challenges of open innovation implementation noted by Chesbrough and Crowther (2006) are potentially heightened for the public sector, and problematized by the regulations around government procurement and interaction with external organisations at all levels, local, regional and national (Mergel, 2017). Furthermore, formal requests for proposals and tendering procedures, with accompanying highly specified product or service descriptions, protect employees and governments from accusations of irregularity or a lack of care with taxpayer funds. Activity in this area is driven by increased consultation with citizens by (frequently city-based) public authorities, and through open government policy initiatives (Bakici, Almirall and Wareham, 2013, Mergel, 2017).

Mergel (2017) reviews the role of the crowdsourcing website Challenge.gov, on which government agencies can post their problem definitions and seek solutions from citizens. The incentive for the vast majority of problems on Challenge.gov was to test the approach, rather than because the agencies involved had no recourse to other solutions or contractors. However, she did discover that agencies who had already experienced using external crowdsourcing sites, such as TopCoder, were more likely to post 'genuine' challenges. Mergel et al (2014) note such interactions between the public and government rarely lead to "*disruptive innovation*"; they

found the most radical innovations were those sought by already cutting edge agencies, such as NASA.

Bakici, Almirall and Wareham (2013) suggest that authorities are evolving from service providers to *"platform managers who run projects and collaborate with third parties and citizens".* They also investigate a mechanism for open innovation in government, but in this case, intermediary organisations, such as Waag Society and the city of Amsterdam. They found that such intermediaries often enact the policies of the councils by reducing the distance between the ideas and the councils via the creation of networks and identification of new ideas. Where the city lacks the capacity to attract or execute the projects themselves, so the intermediaries work (in some cases seamlessly with the city) to do this. These organisations also collaborate on behalf of cities with established sources of potential ideas on behalf of the cities, such as with universities.

Bakici, Almirall and Wareham (2013) anticipate further growth of this kind of collaboration between city halls and public innovation intermediaries in order to engage citizens and civic innovators in the innovation ecosystems of cities and regional governments. However, they also found the city hall usually did not have any strategy or structure for governing these intermediaries. Further barriers to public sector use of open innovation include conflicts of interest, budget limitations, concerns about a lack of accountability of use of taxpayer money, legal and cultural bureaucracy (including procurement rules) and uncertainty about open innovation (Mergel, 2017, Bakici, Almirall and Wareham, 2013).

Mergel (2017) focuses specifically on government's engagement with a *"non-professional"* audience. The relationship between government, open innovation and small to medium enterprises (SMEs) is even more under researched, except on a policy basis. Bakici, Almirall and Wareham's (2013) intermediaries vary in their engagement with SMEs. Certainly, some SMEs will have grown out of initial forays into, for instance, Waag Society's Make or Code labs and events.

One of the most important features of how government facilitates open innovation, both for itself and regarding the wider ecosystem, is expanding the availability of open data. In a policy report, Chesbrough and Vanhaverbeke (2018) state, "*by publishing data in a form that is free, open, and reusable, governments will empower many innovative ideas*."

3.3 Summary

To be able to successfully interrogate how open data is used in practice, it is necessary to define the use that is being investigated – in this case, by innovation in products and services. Innovation can be understood in many different ways. This chapter identifies open innovation as an established model of innovation that allows ideas and knowledge to flow between organisations. I show that there are two parts to a successful open innovation platform - the

architecture (how the innovation is enabled) and the community (where the innovation is sourced from) and outline some common forms of both architecture, such as contests and accelerators, and ecosystems, such as small to medium enterprises and developers. I establish that the public sector engages in open innovation, and that open innovation can encourage a more collaborative approach from the competitive market. The next chapter outlines how open innovation is an appropriate and somewhat established model for use with open data.

Chapter 4 Synthesis: Open Innovation with Open Data

In this chapter I establish that for an organisation to publish or to use open data is to engage in open innovation, from simple outside in processes to more complex coupling. I outline key mechanisms for open innovation with open data, and demonstrate how these have worked in practice. Finally, I define a simple theoretical model of the basic requirements of open data for open innovation, derived from the literature.

4.1 Open Data as Open Innovation

Research in this area is highly emergent. Corrales-Garay et al (2019) note that up to 2017 they cannot find any literature review that covers open data and open innovation in terms of open data constituting an open innovation process. However, they identify research across many areas, including medicine and museology, that discusses some elements of open data and open innovation, or demonstrates a link between the themes, and there is a large body of work on public sector use of open innovation in open government and e-government (Corrales-Garay et al, 2019; Huber, Wainwright and Rentocchini 2018; Mergel et al, 2014; Assar, Boughzala and Thierry, 2011). De Freitas and Dacorso (2014) review the Brazilian Government's compatibility for open innovation, as measured by openness including open data. There are also many studies on the importance of big and linked data for innovation but fewer on open data (Huber, Wainwright and Rentocchini, 2018).

4.1.1 Outbound Open Innovation with Open Data

The release of open data to accelerate innovation from distributed users is based in the strategy of revealing internal resources externally. Such resources are made available without immediate financial compensation, but in anticipation of indirect benefits. Therefore, such activities can be considered non-pecuniary inside out open innovation (Smith and Sandberg, 2018; Corrales-Garay et al, 2019). An example of this is the third pillar of Deutsche Bahn's Mindbox open innovation strategy, where they release open data for anyone to use. The example of Zoopla, mentioned in Chapter 2, is another: on their website they expressly state that they have opened an API to invite, *"developers to create new, exciting applications to distribute this content and change the landscape for those wishing to make more astute property decisions."*[1]

Other researchers see open data as more specifically constituting open innovation architecture. Cohen, Almirall and Chesborough (2017) see open data as an example of the use of a platform to convey the generativity (capacity for novelty) of third parties. Kankanhalli, Zuiderwijk and Kumar

[1] Developer.zoopla.co.uk

Tayi (2017) identify open data platforms as a "*technological mechanism for open innovation*", while Chan (2013) further notes that as the open government license enables reusage, it is therefore part of the open innovation architecture.

Hjalmarsen et al (2014) claim that open data has been the catalyst for much of the growth in open innovation in digital services. Products or services developed in contests with open data provided outbound by organisations, can then be created to be viable within the original organisation which procures or licenses it, or available to the wider market as standalone offerings.

Innovation, acquisition, aggregation, transformation and exploitation of open data by third parties can be considered as part of an open innovation mechanism, rather than something that can be utilised by open innovation. Through this lens, all innovation with open data constitutes open innovation. (Huber, Wainwright and Rentoccini 2018; Smith and Sandberg, 2018; Zimmerman and Pucihar, 2015).

4.1.2 Inbound Open Innovation

Lassinantti (2014) reviewed the European policy documents regarding user innovation and found they were largely focused on user groups in the Information, Communications and Technology (ICT) sector, and emphasized economic benefits. Thus, SMEs and entrepreneurs are a key area of focus. This focus can be seen in European projects supporting open data exploitation such as the Open Data Incubator for Europe. Walker and Simperl (2017) investigate the relationship between entrepreneurs and open data. Analysing the request patterns on data.gov.uk, they find that after private individuals, start-ups and SMEs are the largest categories of requesters. One hundred and nineteen start-ups have made requests for 55 datasets, and 191 SMEs have made requests for 89 datasets since 2014.

They find that not only is open data an enabler of entrepreneurship but also posit that because entrepreneurs engage with open data with the aim of creating sustainable business and a profit motive then they are key for reuse and are keen to overcome the barriers of open innovation reuse that might put off a researcher or developer.

Thorough descriptions of open data users are missing from the literature, and in particular, little is known of who the open government data users are, what motivates them to develop services and how they go about for doing so. There has been considerable research on motivations of comparative innovation groups, such as open source contributors, yet almost none on open data users. Understanding how and when various user groups can be engaged with the various instruments of open innovation is vital. In a 2010 study, Davies finds that users are mostly male, and from smaller businesses in the private sector, the public sector or academia. This chimes with the words of the W3.org paper, 'Best Practice: Establish an Open Data Ecosystem': "*Citizens are not interested in data: they are interested in services being built with the available data and*

information." This suggests that there are only certain phases of open innovation for which citizens should be targeted.

Lassinantti (2014) found that in corresponding topic reports on innovation from various countries, a far more heterogenous view of open data users and innovators was presented. This aligns with the findings of Almirall, Lee and Majchrek (2014) on cities. Devising policy to exclude some of these groups will mean that some users who could be strongly pertinent to the development of impact of open data are excluded without recognition. In failing to acknowledge these groups, and adhering to a myth of open data transformation by the information, communications and technology sector and a rhetoric of data as an economic goldmine, multiple user groups and their role in the open data process are being at best, ignored, or at worst, excluded. European policy, she posits, may actively hinder open data innovation (Lassinantti, 2014).

4.1.3 Coupled Open Innovation with Open Data

There are differing views on exactly what constitutes coupled open innovation with open data. This may be as simple as a third party integrating open and proprietary datasets (Corrales-Garay et al 2019, Huber, Wainwright and Rentoccini, 2018). It may also be where the publisher of the open data benefits from the product or service that is created. Lassinantti's (2014) argument that any open data service is of benefit to citizens, whether directly regarding the public sector or not, can be construed as constituting outside in (and hence coupled) innovation for government.

Given the status of publishing open data as outbound open innovation, and the gathering of ideas as inbound, it is not surprising that inbound and outbound open innovation are found together in the open data context. Ruijer et al (2018) discuss the example of a Dutch rural province, which published open data on their portal then convened a group of civil servants, citizens and students to use the open data to create ideas and insights into the circular economy and healthcare issues. While this certainly comprises outbound open innovation (data publication) and inbound (the new ideas and solutions presented by the external group) it is not clear what the value accruing to the group would be, unless the ideas were successfully implemented. Therefore, it is not true coupled open innovation.

Pharmaceuticals company Boehringer-Ingelheim manages a platform, opnMe, which provides access to pre-clinical compounds and extensive data. This constitutes the outbound open innovation. Scientists can then submit research proposals based on these data and compounds, and successful ones will be developed alongside Boehringer-Ingelheim scientists. This is inbound open innovation, but as it is not clear whether the external scientists can also create value for a new market outside the firm, it is not necessarily coupled open innovation.

A successful example of coupled open innovation with open data is Transport for London (TfL), the public transport authority in London in the United Kingdom. An early open data publisher, TfL originally made forays into app creation. However, these quickly became superseded by private companies such as the early-stage, venture-backed Citymapper. In 2015, over 360 apps used TfL data, growing to over 600 in 2017. TfL, therefore, can focus more resources on the activity of publishing quality open data for use by external organisations. TfL calculate they have a 58:1 return on their investment in Open Data. The app creators also benefit, by selling their products and services to various transport markets, both business to customer and business to business.

Many of the artefacts and processes of the open data ecosystem, including access, repositories and hack activities are essentially indicators that open data culture is moving toward open innovation culture (Temiz et al, 2019). Lee, Hwang and Choi (2012), in a study of public sector open innovation processes across 11 countries worldwide found that "*most countries focus more on the outside-in process of open innovation,*" and "*the inside out process of open innovation has received disproportionately less interest in the public sector*", which aligns with the focus of research described in the previous chapter. Yet, with the advent of open government data, this has largely been turned on its head, as most of the focus is on the publication and dissemination of open data, which constitutes the inside out process.

In the private sector, as discussed, incentives to fully open, and bear the accompanying responsibility and cost, are limited. Companies that have built businesses on open data, such as Open Corporates and Spend Network, do offer their data publicly, but this is cleaned versions of already available data, and as such can be seen as constituting a 'giving back' in line with the philosophy of open data rather than a business decision aimed at creating innovation.

Table 7 Overview of Modes of Open Innovation with Open Data

Inbound (Outside In)	Outbound (Inside Out)	Coupled
Open data acquisition, aggregation, transformation, exploitation activities	Publishing data openly (always present)	Data is published by owner and reused to the benefit of owner (eg TfL)
Less focus, unlike traditional open innovation	OD platforms are a 'technological mechanism for open innovation' Kankanhalli, Zuiderwijk and Kumar Tayi (2017)	Potentially any service of benefit to citizens using OGD
	Primary focus, unlike in the commercial sector (Lee, Huang and Choi, 2012)	Less effective solutions available
	Non-pecuniary IO because of anticipation of indirect benefits (Smith and Sandberg, 2018)	Effective solutions available

4.2 Instruments for Open Innovation with Open Data

Almirall, Lee and Majchzrak (2014), in their study of cities, identify six instruments for open innovation; open data itself, crowdsourcing, hackathons, application development contests, embedded change agents and civic accelerators. Similar activities are identified by Smith and Sandberg (2018) - high quality publication, through the creation of platforms to the organising of challenges to co-creation between cities and developers. While these instruments do not have to be used with open data, or even have data as the focus, they can all utilise it.

The instruments require different types of resources and attract various audiences - citizens, developers, innovation intermediaries and city officials. Crowdsourcing is most often focused on citizens, and civic accelerators generally on SMEs. In some forms of activities, such as urban or living labs, intermediaries who run the labs, businesses, city officials and citizens may all engage. In this way, most open innovation with open data in cities diverges from the established organisation of open innovation ecosystems into competitive markets and collaborative communities, combining them and addressing them as necessary (Almirall, Lee and Majchzrak, 2014).

Digital innovation contests, which include hackathons and application development contests of various natures, are some of the most utilised and flexible instruments that might lead to new and innovative products and services for open government data (Zimmermann and Pucihar, 2015). Competitions to facilitate innovation are a cornerstone of EU Open Data support policy.

Hjalmarsson et al (2014) identify 'digital innovation contests' as the digital evolution of a range of processes designed to enable innovative ideas to be transferred to within a firm for exploitation. They define these contests as "*an event in which third-party developers compete to design and implement the most firm and satisfying service prototype, for a specific purpose, based on open data,*" (Hjalmarsson and Rudmark, 2012). In their study, they include not just hackathons dependent on fully licensed open data compliant with the open definition, but any kind of data shared for innovation purposes.

The key aspect of digital innovation contests is to ensure that the outcomes are aligned with organisational goals. Juell-Skielse et al (2014), in a survey of 33 innovation contests, found that "*organizers design digital innovation contests to function as intermediaries for open data innovation.*" They distinguish five levels of organiser engagement with the outcome of the digital innovation contest.

Table 8 Classification of Levels of Post Contest Support (Juell-Skielse et al, 2014)

Level	Support	Open Innovation
1	No support	Outbound only: third-party developers continue distributed innovation process independently
2	Information and contacts provided	Outbound only: third-party developers continue distributed innovation process independently

Level	Support	Open Innovation
3	Organizer actively offers support to develop an external viable product	Outbound with possibility for inbound: organizer is actively involved in implementation and exploitation
4	Organiser offers internal development support	Intellectual rights either stay with third party (inbound continues) or are transferred to organiser
5	Organiser incorporates developer	Open innovation becomes absorbed into internal innovation process

It is challenging to measure innovation with open data through digital innovation contests because they do not follow existing innovation value chain patterns (Ayele et al, 2015). As can be seen from Juell-Skielse et al's (2014) levels of support, there is potential for involvement of the organiser at the deployment end of the value chain. Ayele et al (2015) attempt to measure innovation with open data by creating a measurement model for digital innovation contests, by adding new dimensions to existing innovation measurement frameworks.

They review 13 different models of measurement of innovation and attempt to define one that will work for innovation contests, the key issue being the different levels of support organisers give. Their model assesses input, activities and output and measures at three different points - planning, ideation and service design. In particular, they develop a problem – solution maturity index inspired by Mankins (1995). This index measures how defined the problem is and how effective known solutions are in solving the problem. This index is intended to assist digital innovation contest organizers to formulate problems and to evaluate ideas and solutions.

Table 9 Problem-Solution Maturity Index for Digital Innovation Contests (Ayele et al, 2015)

Maturity	Problem	Solutions
Low	Unspecified	Lacking
Medium	Specified	Lacking
High	Specified and acknowledged	Less effective solutions available
Very High	Clearly specified and highly acknowledged	Effective solutions available

Accelerators focusing on open data are less well researched. The Open Data Institute[2] has run start up incubators – collaborative programmes designed to help firms succeed - for companies acquiring, aggregating, exploiting and transforming open data for a number of years. Civic accelerators are a subset of this kind of activity, run by city authorities. The Guelph Civic Accelerator is a Canadian initiative that enables the city of Guelph to engage in open innovation with entrepreneurs, start ups and students to create solutions for public sector problems. It comprises an innovation challenge, inviting proposals for solutions to specified and acknowledged problems with less effective solutions available (Ayele et al's 'High' category). The winners enter an accelerator programme, working with city staff to develop their pilot solution. They are given the opportunity to win a contract with the city at the end of the civic accelerator.

[2] www.theodi.org

4.3 Impact of Open Innovation with Open Data

The Apps for Democracy contest in 2009 was launched by the government of Washington DC to support use of their recently opened datasets. The contest cost $50,000 and, according to its creators, returned 47 iPhone, Facebook and web applications with an estimated value in excess of $2,300,000 to the city. Inspired by this success, cities with recently opened data catalogues worldwide began hosting application contests (Lee, Almirall and Wareham, 2016). These contests continue to be the dominant strategy for fostering transparency and economic development provided by civic open data. Such promotion of re-use may be run by governments, such as Stockholm Open Lab, start-up groups, such as Belgian co-working space Betacowork's Open Data Hackathon; by a combination of academia, activists and government, such as Apps for Ghent; or by corporates, such as German transport operator Deutsche Bahn, which runs regular hackathons with its open data in Berlin.

These events can be the catalyst for people and ideas to come together to create a firm. Energy saving big data processing firm Mastodon C was created after an initial open data hackathon weekend. However, such stories are the exception, and these initiatives have suffered from a lack of impact, both within government and the public. Hackathons manifest an alignment with the Western, neo-liberal view of entrepreneurship, most often seen through judgements on the success of hackathons deriving from assessments of potential size of markets, or estimated financial Gross Value Added (Irani, 2015). This is not intrinsic, nor have hackathons historically been proven fora for the sustained development of start up businesses (Davies, 2013, Irani, 2015). Hivon and Titah (2015) note that hackathons are often used for skills acquisition and are therefore as much about knowledge transfer as innovating. Porway (2013) notes that "*data hackathons often lack clear problem definitions"* and also notes that they solve the participants' problems from their individual point of view, but not necessarily the larger challenges they are intended to address. Few of the innovations in inside out open innovation contests end up as viable business propositions (Hjalmarsson et al, 2014; Lee, Almirall and Wareham, 2016; Chan, 2013). This may be as low as 10% (Hjarlmarsson et al, 2014).

While data quality and general interest were initially seen as the locus of the failures, a variety of other reasons have emerged, some to do with the contests themselves, but many to do with pre- and post-contest engagement. One risk to do with the design of the contests is a low bar for participation. Where the requirement is the simple inclusion of a city- provided open dataset, many developers submit previously developed apps with minor adjustments to accommodate the key datasets (Lee, Almirall and Wareham, 2016).

Lee, Almirall and Wareham (2016) identify three key areas for improving the success of innovation with open data. Firstly, limited public knowledge of the operational challenges facing city governments means that apps are often not targeted in areas of most value. Frequently they

address a consumer audience as it is easier for the developer to understand that space, than the opaque and highly regulated civic arena. Public administrations are strongly positioned to address this, but this requires increased exposure of innovators to civic needs in open data challenges. Secondly, stronger management of open data initiatives within the authority is needed. This would address issues such as that which Lassinantti (2013) found in Stockholm. Lastly, 'common platforms for open data initiatives', or standards, play an important role, particularly in the possibility of replication or uptake in multiple cities. This increases the opportunity for a new product or service to be utilised, but also reduces the cost of innovation for cities.

4.4 Barriers to Open Innovation with Open Data

Lassinantti (2013) notes that while the external data users in the cases she studied were consistently actively involved, the public sector participation varied in terms of how active it was. In a study of open innovation practices in Stockholm and Skelleftea she found that the city of Stockholm's open data strategy was directed by the IT department, which acknowledged the existence of developed apps, but, seeing themselves as providers of data rather than part of the open innovation value chain, did not enter into discussions or collaborations with the service developers. As a result, the value chain was broken.

The nature of open data, meaning there are no Service Level Agreements, is problematic - there are no assurances in open data publishing (Smith and Sandberg, 2018). The risk in innovating with open data is of course that the data disappears or degrades. Huber, Wainwright and Rentoccini (2018) note the example of a SME which experienced product failure when a key source of open data was closed. While this can happen under other circumstances, there is no warning or redress with open data. This can be mitigated by a closer relationship to the original source. However, Smith and Sandberg (2018) identify 38 barriers associated with open innovation with open data and associate them with five phases of service development and found that across four of the phases the issues were mostly to do with relationships with the open data publisher, whether this was because they were not publishing sufficient data, or not communicating well, or not giving sufficient direction. This is particularly problematic for SMEs, which are often unable to take advantage of open data because of a lack of available resources and difficulties in forming external partnerships. In terms of open innovation, the ability to engage with open data publishers is particularly important for the acquisition and assimilation capability. (Huber, Wainwright and Rentoccini, 2018; Walker, 2014).

The other barriers and benefits they describe are largely the same as the general barriers and benefits of open data. Part of this is relying on agencies to correctly identify datasets of potential value to user communities. This requires robust internal mechanisms for engaging the right parties in preparing and approving datasets for release (Kankanhalli, Zuiderwijk and Kumar Tayi,

2017). Demeyer et al (2012) describe their experience of running a hackathon in Amsterdam, noting that despite having few stipulations about the format and quality of the data they sought, *"It turned out that finding new and relevant data was the hardest part of setting up the contest. It took many days of efforts from all organizing parties to gradually open up more datasets.*"

While hackathons are the dominant form of Digital Innovation Contest, the intrinsic format - they usually take place over a full weekend – can be a problem. Such intensity imposes a cost of entry requirement that excludes those who perform paid work at weekends such as those in retail or public transport, those who have childcare or family commitments or even those for whom spending an entire weekend in the company of strangers is problematic. The UK Accountability Hack involves participants sleeping at the National Audit Office. This extends the productive period but clearly also designates the type of person who might be expected to attend. Gomez-Cruz and Thornham (2016) argue hackathons also embody clear age, gender, class and ethnicity discrimination. This is in part down to the values implicit in choices made by organisers, but also in that the practical requirements for participants include the possession of certain skills and often, possession of a laptop.

Critics argue that the data considered useful in hackathons (clean, processed, big) enacts a discriminatory politics. Gomez-Cruz and Thornton's (2016) ethnographic study of 20 hackathons found that even where specific people from a variety of demographics were invited to participate in hackathons, they could not identify with, nor recognise themselves within, those datasets.

Hackathons require people to be able to code in multiple languages, be familiar with Application Programming Interfaces (APIs), and more. As well as coding skills, participants may need to be proficient in fuzzy matching, thinking in graphs (that is, in terms of nodes and edges), require familiarity with code repositories and also to be able to switch coding languages dependent on features of the API (Walker, 2017). Hackathons to an extent exclude end users, who matter less than the developers (Irani, 2015). Engaging with theoretical and technical challenges in designing for true end-user use is not inherent in the hackathon (Dix, 2014).

In terms of ability to create novel products and replace proprietary data with free data, boundary spanning is important in both open data and open innovation, requiring organisations to have strength in data science and in marketing the new products. Transformation barriers were found mainly in recruiting skilled data scientists (Huber, Wainwright and Rentoccini, 2018). Based on a three-month study of a Swedish digital innovation contest (Travelhack 2013), Hjalmarsson et al (2014) found that a lack of time and money to develop the initiative after the initial contest was the greatest barrier to innovation. They found this was particularly true where the competitors were students or researchers, which is plausible, given that policy focus for support is aimed at SMEs, rather than either of these two groups. The challenges of open data hackathons are compounded in the open innovation model if there is no overarching organisation interested in

developing and distributing any 'inventions' emerging from these hacks, nor a channel for sharing the 'excess' potential innovations with other entrepreneurs or organisations.

Finally, there are barriers around the open data itself. Huber, Wainwright and Rentoccini, (2018) describe open data as *'fuzzy*' in practice, especially where it is not fully open (for instance, published in a difficult to use format, or without clear licensing).

4.5 A Framework of Open Data for Open Innovation

To explore whether open innovation with open data is being practised in conformance with the literature it is necessary to define a framework that encompasses the key issues of the literature. In this section I present such a framework. This constitutes the indivisible requirements for open innovation with open data, as represented in the literature. Open data is both a very simple premise, and, as the previous chapters have shown, extremely challenging. While many other features could be added, these are 'nice to haves' rather than the 'must haves' captured in this framework - the parameters that must exist to fulfil both the data and innovation aspects. This aims to be a reductive framework. Unlike the Common Assessment Framework or other open data model this does not seek to assess open data in any way or judge a specific standard, quality or outcome that is must achieve. It simply aims to capture the minimum requirements for open data in the context of open innovation.

This framework is based on three aspects:

1. Features expressed in the simplest open definition;
2. Elements created by law, primarily copyright and the GDPR;
3. Features necessitated by the requirement for innovation.

Some of these features - purpose and access - are derived directly from the Open Definition. This is also in line with Longshore Smith and Seward's (2017) praxis definition of open as 'anyone can participate'. Although they also suggest that to be open the artefact should also be free, most definitions of open data suggest the possibility of marginal cost, so this is not included here.

There are two legal elements that are absolutely core to the framework. The first is the derived from the license, and therefore indirectly from copyright law. The license – in whatever version it is – is the mechanism by which the data is identified as specifically open data. This grants the user permission to use the data. The second is derived from data protection laws. Open data cannot conflict with the GDPR, so this is an important boundary for the framework.

To set the framework in the innovation context, the literature equates open data with open innovation. In other words, for an organisation to publish or to use open data is to engage in open innovation (Smith and Sandberg, 2018; Corrales-Garay et al, 2019; Cohen, Almirall and

Chesborough, 2017; Kankanhalli, Zuiderwijk and Kumar Tayi, 2017; Chan, 2013). Organisations can further engage in more complex open innovation by using other mechanisms such as civic accelerators or digital innovation contests with open data, or a combination of many possible mechanisms, but in its simplest form, publishing open data is outbound open innovation, and using open data is inbound. However, explicit value must be added to the framework in order to drive investment processes and value creation. Without the benefit that will be gained from open data there is no imperative to open.

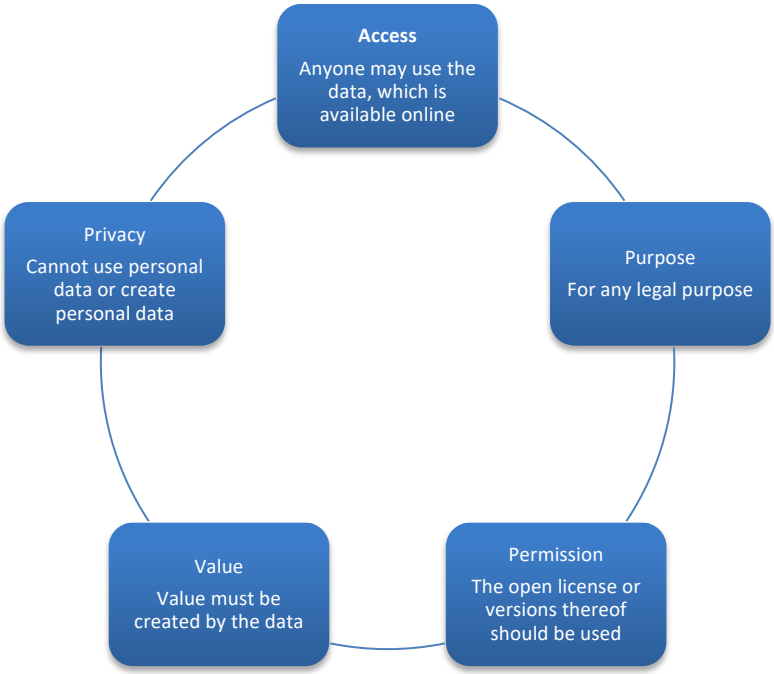

Figure 4 Framework of Open Data for Open Innovation

The features can be understood as follows:

Purpose: Open data is agnostic as to use. This is perceived as a strength for innovation, as it does not curtail any use nor focus invention in a specific area.

Access: Access is equally available to all. Provided people are skilled enough to discover the data and then utilise it, they can use it. Multiple programmes exist to support this. Again, this is seen as a positive aspect for innovation, as it removes first mover advantage that might accrue to larger firms with greater resources.

Permission: Although mention of a license is not explicit in the open definition, data without an open license confirmation is simply publicly available information, so this is a core requirement. Again, this is a strength for innovation as often, (although as seen in Chapter 2, not always) there are no rights-based complexities associated with this.

Privacy: It is intrinsic to establishing a basis of comfort with the idea of public sector information being made available for use (including innovation) that open data is not personal data. Privacy is therefore a boundary of open data - once this is crossed, it cannot be open data.

Value: While value is not intrinsic to open data itself, it is the key motivator, for opening and use. The complexity is in defining where the value lies. Often the value is indirect (largely the value is indirect through tax payer benefits), or, per Lassinantti (2014), widely distributed. It is often difficult to identify where the locus of the value is in open innovation with open data because unless a coupled open innovation mechanism is used, tracking is not possible, and there are few impact models.

4.6 Summary

This chapter establishes that for an organisation to publish or to use open data is to engage in open innovation (and consequently, all innovation with open data is open innovation of some kind). Publishers, especially at the city level, are interested in engaging with entrepreneurial and start up companies to develop open innovation solutions to public sector problems. Limited knowledge of the operational challenges facing city governments means that apps and solutions created by SMEs, developers and others are often not targeted in areas of most value, and consequently there is little post-creation support for them. Public administrations are strongly positioned to address this, and can do this through digital innovation contests and civic accelerators.

The three background chapters have covered the existing relevant literature on open data, open innovation, and the theoretical model of open innovation with open data. Based on the literature, a simple model has been devised that outlines the 'rules' of open data for open innovation as it is currently conceptualised. This answers the first of my research questions, 'What are the key components of a theoretical model of open innovation with open data?' enabling comparison of the praxis of open innovation with open data with this model.

The following chapter presents the case study of public sector open innovation with open data, where city authorities engage with SMEs via digital innovation contests and civic accelerators, to develop new services with their open data.

Chapter 5 Case Study: Smart Cities Innovation Framework Implementation

The Smart Cities Innovation Framework Implementation (SCIFI) is an open innovation project that brings together cities with a public service need with SMEs which aim to create a solution in part based on the open data of the cities. It is a technological and social innovation project in response to Interreg 2Seas European Regional Development Fund (ERDF) programme priority specific objective, *"improve the framework conditions for the delivery of innovation, in relation to smart specialisation"*. Covering coastal areas of England, France, Belgium (Flanders) and the Netherlands, the 2Seas area is connected by the titular two seas of the Channel and the North Sea.

The market for public sector open data in Europe is €52bn (European Commission, 2018). Regional Smart Specialisation Strategies (RIS3) encourage smart innovation. For instance, cities are looking for parking solutions for their congested centres, and the private sector has the skills and technologies to facilitate the development of these. However, the 2Seas innovation sector faces shared public sector challenges in unlocking this value for businesses. The market failure has been previously identified as being mainly within 2Seas cities in opening up datasets and using innovative procurement. The partners within the SCIFI consortium (4 cities, a university, a living lab and 2 business networks) desire to identify transnational framework conditions that create a cross-border market that gathers the fragmented resources of cities and innovation potential of businesses, and increases the capacity of cities to activate the dormant demand for smart public services.

To activate demand, the cities aim to increase their technical capacity in open data and innovative market engagement in order to connect supply and expertise. Innovators are connected through the two business associations and transregional sister organisations. The living lab and the university leverage their expert role in transnational open data initiatives. (These include Open Data Incubator for Europe, Open and Agile Smart Cities, FIWARE and the European Network of Living Labs.) To demonstrate value SCIFI uses a digital innovation contest and civic accelerator programme to test the framework conditions for innovation in mobility, energy, and clean environment. The living lab and university coach cities and businesses in challenge identification, procurement, solution co-creation and implementation. Success stories are intended to demonstrate the value from open data.

The initial activities within SCIFI map to the elements of digital innovation contests as outlined by Hjalmarsson et al (2014). The cities, along with citizens, business and academia, devise challenges with associated data, which are then promoted to the SME community across Europe, via a call mechanism. Applications outlining suggested solution approaches are received and filtered and interviews conducted with the most promising applicants. Once the SMEs have been chosen to

partner with the cities, they co-develop solutions in a six-month civic accelerator, working closely with the cities and other consortium partners. At the end of the accelerator, the intention is for a procurement opportunity for the new product or service, either with the developing city or another city with similar issues. In some instances, the solution is being developed in more than one city, or different solutions to the same challenge are being developed in the same city. At the end of the civic accelerator, successful pilots are procured in a separate process.

Table 10 Elements of SCIFI Digital Innovation Contest(Based on Hjalmarsen et al, 2014)

Design Elements	Attributes	
Needs: financial and municipal resources to develop prototypes meeting user (city departments, citizens) needs	Resource: provide details of smart city need	Facilitation: support teams interpreting user needs (city departments, citizens)
Value: financial and networking resources to develop prototypes generating value	Resource: provide living labs support and transregional promotion to other smart cities	Facilitation: support teams with business value issues
Data: provision of open data addressing the contest space	Resource: present available open data in an engaging way	Facilitation: support teams with API issues (mainly required JSON)
Novelty: input stimulating teams to ensure novelty in output	Define rules for intellectual property (contract)	Provide baseline for innovation (in plan, not achieved)

Table 11 SCIFI Project Timeline

Date	Activity
July 2017	Project Launch
October 2017	Steering Committee 1
January - June 2018	Development of business cases (challenges)
March 2018	Steering Committee 2
June 2018	Steering Committee 3
July 2018	Launch of digital innovation contest (Call 1)
October 2018	Steering Committee 4
January 2019	Launch of 1st civic accelerator (8 pilots)
January 2019	Steering Committee 5
January - June 2019	Development of business cases (challenges)
April 2019	Steering Committee 6
July 2019	Launch of digital innovation contest (Call 2)
July 2019	End of 1st civic accelerator
October 2019	Steering Committee 7

Date	Activity
January 2020	Launch of 2nd civic accelerator
January 2020	Steering Committee 8 Procurement process for successful pilots begins
June 2021	Project ends

5.1.1 The Smart City Context

Much of the research on digital innovation contests for open innovation has been conducted in city contexts. This is largely because of the relationship between open data and smart cities. The city is a particularly important context for open data, which is seen as a "*defining element*" of smart cities (Ojo, Curry, and Zeleti, 2015).

5.1.1.1 What is a Smart City?

The term 'smart city' covers many concepts. At the most implemented end is the ubiquitous 'smartcard' such as Oyster in the UK, OV-ChipKaart in the Netherlands or LIsboaViva in Portugal. At the most conceptual, advanced, end of the spectrum lies the (recently abandoned) Sidewalk Labs and Toronto City 'Quayside' project, with autonomous cars, intelligent rubbish collection, smart air quality measurement and heated streets. Essentially, it is a way to use technology and data to improve life for citizens.

In their metastudy of smart city definitions, Nam and Pardo (2011) outline a smart city as constituting a Venn diagram of technology, human and institutional dimensions. Recent models, my own amongst them, centre on people, but suggest data is the next most crucial ingredient (Walker, Ibanez and Simperl, 2019).

Figure 5 Citizen-centric Factors of Smart Cities
(Walker, Ibanez and Simperl, 2019)

5.1.1.2 Smart Cities and Open Data

In the United States alone, more than 40 cities have open data portals for use by citizens and private firms (Cohen, Almirall and Chesbrough, 2017). In Europe, The Urban Data Platform project aims to ensure that 300 million European citizens are served by cities with open data platforms by 2025. Cities have embraced sensor data, which, with other Internet of Things data, will shortly become the pre-eminent form of data. The smart city context enriches the open data ecosystem, and shapes open data initiatives. In their study of 18 open data initiatives across 5 smart cities Ojo, Curry and Zeleti (2015) find that initiatives around 'innovation cluster data' such as transport and mobility have more focus, which supports their notion that a data-oriented city is an open innovation economy.

5.1.1.3 Smart Cities and Open Innovation

Cohen, Almirall and Chesbrough (2017) state that, "*the future of innovation will require collaboration and co-creation with local governments*" and this will rely deeply on the availability of data. Open data may be the best transition so far of the platform model to Smart Cities, however, they note that the business models to drive the effective use of open data in smart cities are still somewhat lacking. As much as collaboration with the public administration is the way forward, municipalities are not necessarily skilled in managing open innovation: they lack the skills and processes for engaging with third parties, and often suffer from integration and conflict issues between third party and internal applications (Cohen, Almirall and Chesbrough, 2017). They note the example of Amsterdam Smart City, which has an external agency more skilled in developing such partnerships, but as a result of being external, the agency lacks the decision-making capability and influence.

City Background and Open Data Experience

This is a single case study, that is, of the Smart Cities Implementation Framework Initiative open innovation with open data to create 'smart city ' style public services with data and technology. There are 4 cities within this, who are the initiators of the open innovation and the publishers of the open data. Two main reasons mean that there is not much insight to be gained from inter-city comparison. The first is that the cities are constrained by the framework to work within similar formats, removing some elements of choice and artificially creating similarities of approach - each city is being supported by the same knowledge and network partners within the project. The second is that although the cities are all in their early stages of the open data journey, there is sufficient difference in expertise, history and focus to reduce the value of comparison. Consequently, the case study should be seen as a single unit. However, details about each city are important in order to understand the overall background of the case study.

The following are based on a programme deliverable (WP1_D1.4.1_Questionnaire). This is a questionnaire regarding the status of the cities vis-a-vis open data developed by one of the cities at the beginning of the programme.

5.1.1.4 City 1

City 1 is situated in Flanders in Belgium, half way between Brussels and Antwerp. It has a population of 86,000.

In addition to the Flemish decree on the reuse of government information (April 2007, adapted in 2015), there is a specific Flemish Open Data Policy Framework. On the general disclaimer of the public website it is mentioned that the city retains all intellectual property rights on the website and all information published through the website. The GDPR is translated into national legislation, which City 1 feels can cause additional restriction.

There is no formal definition of open data within the city organisation. Open data is regarded as any data that is published (or will be published) by the city organisation in a structured way on their/a website. In a narrower interpretation, it is limited to structured data (not including pdf documents or images/videos). There are some references to an information strategy in the general Policy Note 2013-2018 adopted at the start of the current city council. The city has no open data policy, but will (probably) sign and subsequently adopt the Open Data Charter as created in the Smart Flanders programme. The Open Data Charter is seen as a document that is useful as a tool to introduce and maintain the necessary principles for open data in the city. It is hoped it will be the driver for a cultural shift within the organisation towards transparency, innovation and data-driven policies.

The city has an ambition to be exemplary for an open authority. To be a Smart City is defined as a organisation-wide goal, thereby implicitly supporting the open data publication. City 1 has a smart city vision and strategy document. Their conceptual target audience for open data is anyone interested in creating added value for the city and its stakeholders (the "*common interest*") although communications are limited to the occasional Smart City related press release or article for policy or data reuse.

At the beginning of the project, City 1 had yet to publish any open data.

5.1.1.5 City 2

City 2 is a large city of the province of West Flanders in Belgium, in the northwest of the country, and the seventh largest city of the country by population. It has a population of 117,000, and is visited by 8.3m tourists a year. As it is in Flanders it is subject to the Flemish Open Data Policy Framework and GDPR as above. The city defines open data as any data that the city gathers in view of its public role. They do not have an information strategy but they have an approved Open

Data Policy since August 2017. The policy is aimed to serve as reassurance for city-services and policymakers hesitant to publish Open Data.

The city has a website that published, at the beginning of the project, 12 data sets. It has a free open data license. The target groups for their open data are companies and knowledge institutions, and they held a hackathon. Other than this, they have not had a policy of encouraging data use, and have no metrics for output. As well as stimulating innovation, they see opening data as leverage to improve their own data for internal use.

5.1.1.6 City 3

City 3 is a French city of 56,000 in northern Hauts-de-France, south of Lille. French regulation obliges cities with more 3,500 inhabitants to publish their data, on the principle that citizens have the right to access all databases of a public administration. The city publishes their internal datasets only if they meet at least 3 of the 5 stars of open linked data in that they are in an open format and machine readable. There is restriction on data relating to natural persons and also data related to security, but there are no restrictions on use or intellectual property rights.

Open data is a part of a wider smart city strategy. They aim to reuse data to improve municipal services by offering new services to citizens or increase the efficiency of internal services. At the start of the project open data was used to Inform internal departments and citizens about main indicators of the local territory. Their main targets for their open data are companies and knowledge institutions. They communicate about open data with other stakeholders mainly through data visualisation.

The city had very little data available at the beginning of the project.

5.1.1.7 City 4

City 4 has a population of 101,000 and is situated in South Holland in the Netherlands, between Rotterdam and Den Haag. It is the home of the largest technical university in the Netherlands, which has also been a leader in open data research over the past decade.

There are a number of national mandates for publishing open data but the primary one is the 'Wet Openbaarheid van Bestuur' (Law on Open Government). It states the rights of citizens to know what the government does, which choices it makes and how money is spent, with the goal of creating an open government. Currently the law is somewhat reactive, however a new law is in the making that is more proactive.

The city has had an open data strategy since 2016. The aim is to create value - transparency, improved quality of data, public services and direct and indirect economic value - from the data generated and collected by public services. The city's open data strategy is related to its Smart City strategy. In this, the city aims at becoming 'open by design,' by using open data as an instrument and standard in new projects and developments. There are four underlying principles:

(1) they publish data based on political and organizational priorities, (2) data is only published if the organizational, judicial and financial concepts are clear, (3) only in cooperation with internal and external stakeholders can the city create value with open data and (4) they have clear rules on technical implementation and regulation regarding information/data standards.

Their target audience is “everyone”, but specifically developers and citizens. They promote their open data policy through speeches, newsletters and social media, and promote use with some social media. So far, they have been unable to formulate key performance indicators for their open data publication and use as this is too complicated. They are developing internal performance indicators, such as how they process incoming demand for open data.

Table 12 City Experience with Open Data

	City 1	City 2	City 3	City 4
Open Data Strategy	No	Open Data Policy	Part of Smart City Strategy	Yes, since 2016
Open Data Sets 2017	None	12	Very few	17
Target Audience	Anyone	Companies and knowledge institutions	Companies and knowledge institutions	Developers and citizens

From the above it can be seen that all the cities are in the earliest stages of their open data journeys, although this ranges from no experience at all, to some limited experience.

5.2 Summary

The SCIFI project makes an appropriate case study as it is an opportunity to study a clearly specified open innovation programme using open government data from a number of city authorities. With 4 different locations under one umbrella there is ample redundancy, reducing the risk of a single iteration failing and therefore being unable to collect data. It has clearly specified mechanisms, the digital innovation contest and the civic accelerator, and a target aim of the competitive market, specifically SMEs. In the next chapter, I describe my research methodology.

Chapter 6 Methodology

The methodology comprises two approaches: a case study to address research question 2 (RQ2), and an integrative literature review to address research question 3 (RQ3). For RQ2, the case study research design was selected as it is appropriate to the research question and the topic. Corrales-Garay et al (2019,) in their overview of the literature on open data and open innovation, found that of the empirical (as opposed to theoretical) studies, over 60% used the case study approach. Of previous literature reviewed in this thesis, Juell-Skilse et al (2014) use the case study strategy when devising their framework for measuring open data innovation, as do Bakici, Almirall and Wareham (2013) in their analysis of the role of intermediaries in public sector open innovation.

The case study approach enables a detailed description of a contemporary phenomenon within its context (Yin, 1994) and is particularly suitable for examining whether theory is applied in practice (Gerring, 2004). Document analysis and group interview were the methods employed for data collection, and thematic and content analysis were utilised.

For RQ3, the integrative literature review is "*a distinctive form of research that generates new knowledge about the topic reviewed"* (Toccaro, 2005). Similarly, it was chosen for applicability to the research question and because this is an emergent area of study. It has been used in relevant and related disciplines such as information systems, management and social science (Toccaro, 2016).

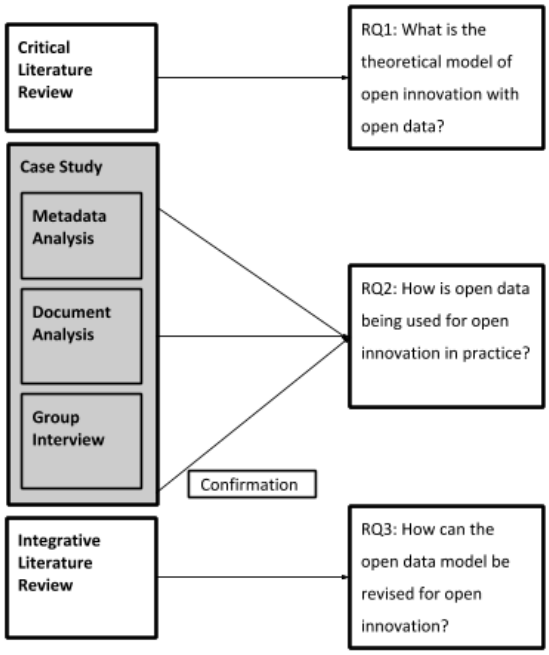

Figure 6 Outline of the Research Methodology

6.1 Structure of this Methodology

First of all, I address RQ2. I begin by discussing the case study research strategy. I then move on to describe the specific case study setting that will be used in this particular research, and then establish key background about the context (smart cities) and the operational setting (the Smart Cities Innovation Framework Implementation project). Subsequently, I outline the data gathering and data analysis approaches used, before going on to describe these in depth. Finally, I address RQ3's research approach, the integrative literature review, in depth.

6.2 Research Question 2: Case Study

6.2.1 Defining a Case Study

First of all, a disambiguation. This is a research case study, as opposed to a pedagogical case study of the type frequently used in business school teaching which aims to *"elucidate features specific to a particular case"* (Gerring and Seawright, 2008). Even within this narrower field, the definition of a case study can be somewhat ambiguous.

The case study is a way to study phenomena within their context (Hartley, 2004; Yin 1994; Gibbert, Ruigrok and Wicki, 2008). It allows for intensive focus and detailed investigation of the phenomena and context (Gerring and Seawright, 2008; Hartley, 2004). Therefore, it is of great use where the boundaries between phenomenon and context are indistinct and where dynamics play a part (Yin, 2003). The unit of investigation can be single or multiple, with the goal of understanding a larger population of similar cases (Gerring and Seawright; Hyett, Kenny and Dickson-Swift, 2014). Yin (2003) argues that findings cannot be generalizable to populations but they are they are useful for generalizing theories. Stake (2005) argues that what is of interest is what can be learnt from a single case about that case. Both of these are applicable in this situation.

"By whatever methods, we choose to study the case" (Stake, 2005). An important feature is that it is a research strategy rather than a method (Eisendhardt, 1989; Hartley, 2004; Hyett, 2014). It is defined by the focus on individual cases rather than by the methods of inquiry utilised, which renders it flexible (Hyett, 2014).

6.2.2 The Purpose of a Case Study

Case studies are useful for generating hypotheses and building theory from data (Eisenhardt, 1989; Hartley, 2004). Development of a case study benefits from the "*prior development of theoretical propositions to guide data collection and analysis,*" (Yin, 2003). This fits the purpose of

this research. Further, case study research has frequently been used in examining innovation in organisations (for example, Leonard-Barton,1987; Dougherty and Hardy, 1996).

6.2.3 Critiques of the Case Study as a Research Strategy

Utilising case studies is often seen as very complex and challenging (Yin, 2003). Consequently, it is sometimes critiqued as lacking scientific rigour. Researcher bias is another concern, although this extends to virtually all qualitative methods. Multiple researchers are often recommended to avoid this (Yin, 2003; Eisenhardt, 1989; Miles and Huberman, 1994). However, the majority of concerns around case study use concern the extent to which it is possible for researchers, practitioners or policy makers to interpret the data consistently across individuals and arrive at the same conclusions as the initial investigator (Sinkovics, 2018). These concerns often arise as epistemological disputes.

6.2.4 Epistemological Differences

Yin's (1994) work on conducting case studies is possibly the most widely cited guide. This takes a strongly positivist approach. The positivist tradition in qualitative research attempts to translate some of the operational aspects of quantitative research - for instance, reliability and representativeness - into the qualitative setting (Lin, 1998). Positivists locate validity in the extent to which the results of one research study can be replicated in another study. Subsequent work by Yin (2012) has taken on a post-positivist hue, which acknowledges the impossibility of perfect objective reality in human knowledge.

A key alternative epistemological approach, of which the foremost champion is Stake (e.g. 2005) is interpretivism. Interpretivists locate validity in credibility and accuracy of description. The method of inquiry is more transactional and locates the researcher within the research (Hyett et al, 2014).

Lin (2005) argues that a positivist or interpretivist approach alone is insufficient to explain both the 'how' and the 'why' of causality. Positivist work identifies general patterns, while interpretivist work illuminates how that pattern works in practice, and consequently, the two must co-exist (Lin, 2005). Similarly, Hyett et al (2014) argue for post-positivist rigour, while also acknowledging that the value may be more understood through an interpretivist (or social constructivist) lens.

As Lin (1998) notes, the accurate identification of a causal mechanism between two (important) variables not only enhances the "*causal story*" but changes it. As an example, if local government is not opening data, it matters whether that reason is because it is not mandatory, or they have insufficient IT resources, or because of a possible chilling effect of GDPR (or all three).

The case study approach therefore offers flexibility between these two paradigms (or as Luck et al (2006) have it, a bridge between the two). Such a combination of theoretical approaches is not

novel (Hyett et al, 2014), and can be facilitated by the use of rigorous and replicable thematic analysis with an analytical codebook (Fereday and Muir-Cochrane, 2006).

This case study utilises predominantly interpretivist approaches the research gathering and focus, using documents and group interview to understand the experience of a specific incidence, however, the analysis of the research data is underpinned positivistly, with a replicable approach using thematic analysis and a codebook initiated with deductive coding.

6.2.5 Selecting the Case Study

Case study research is an interdisciplinary practice, which means a clear methodology is more important that an approach that is based in a single discipline and relies on existing assumptions (Hyett et al, 2014). The key decisions in designing a case study are: defining the case of study; determining the method of data collection; and analysing and presenting the data (Yin, 2011).

Case selection and case analysis are commonly intertwined more than in other research methods (Seawright and Gerring, 2008). They suggest the case study is called on to perform a "*heroic role*", standing as representative for other, non-investigated cases, arguing while it cannot stand for everything, it must stand for something. Stake (2005) argues against the possibility of representativeness of anything but the case itself. Both agree, however, that case selection must be theoretically driven and not random. The other aspect of selection is the number of units that comprise the case study. Stake (2005) identifies three types of case studies: the intrinsic (which is used to understand the details of a single case); the instrumental, which provides insight on an issue or is used to refine theory; and the collective. Only the collective requires multiple units, as it is an instrumental case which is studied as multiple cases (part of the same case study).

The case here was selected for its typicality and its suitability as an instrumental case. While this was not, as Seawright and Gerring (2008) suggest, identified by a statistical test, it was identified from my experience working on multiple open data and European projects, including the Horizon 2020 open innovation data sharing project Data Pitch, and the European Data Portal. While there was an element of pragmatism that all researchers require, in terms of time, access, resources and expertise, it is important that it is understood how the particular case study fits within the pantheon of possible case studies and therefore, what it contains and what it represents. Below I introduce the case study setting and its theoretical underpinnings.

6.2.6 Justification for Selection

There is typicality here, as suggested by Stake (2005). The public sector units are cities, rather than agencies; this reflects other research, by, for instance, Lassinantti (2014) and Bakici, Almirall and Wareham (2013). The consortium and its aims also have counterparts in other regions of Europe (for instance, Interreg VB project Smart Cities and Open data REuse (SCORE) or Interreg

NW Europe Building an Ecosystem to Generate Opportunities in Open Data (BEGOOD)). They are also representative of medium-sized northern European cities in many ways. However, there is also opportunity to explore new areas, as the cities are combined together in a consortium, which has been identified as a gap in the open innovation knowledge (West and Bogers, 2013). They are not developing apps in a contest, but rather services and products that they themselves wish to utilise.

Further, the participants in the SCIFI project and the documents are accessible to me as a researcher, as the University of Southampton is a 'knowledge partner' in the SCIFI consortium.

6.2.7 Limitations:

This is a single case study and as such is subject to particular effects which may not repeat elsewhere. External validity, for the extrapolation of theory, would be increased by the exploration, comparison and synthesis of similar projects for instance, Interreg VB project Smart Cities and Open data REuse (SCORE) or Interreg NW Europe Building an Ecosystem to Generate Opportunities in Open Data (BEGOOD). However, there are 4 municipalities involved in the single case study, so this may assist in guarding against the situation to be too particular or individual to be of relevance.

6.3 Case Study Data 1: Document Analysis

As previously noted, although case studies embody a useful descriptive methodology there is no one agreed method for execution, either in terms of identifying and collecting the source material, or the analysis of the data. A key concern is that the methods employed are performed in a way that is transparent and credible. It is fully possible to use quantitative and mixed methods within the work, (Stake 2005) although the success of this depends on what is being analysed and the number of instances.

Two forms of data gathering are used; document analysis and group interview. The document analysis is split into two parts; analysis of written texts and analysis of the many lists of data used in the project and pilots (metadata). While this is not genuine triangulation, as defined by Denzin (2006), it does allow written statements about data use, availability or other dimensions to be compared against the information in the metadatasets themselves. The results of the two analyses are presented to the cities and other members of the consortium in a group interview, for feedback and comment. Again, this is a confirmatory interview rather than triangulation, but it allows for input regarding the researcher's interpretation and accuracy.

6.3.1 Document Analysis

Although participant interviews are seen as a gold standard, the information gathered is based on the interviewee's account of actions that occurred elsewhere, both in terms of time and place (Becker and Gear, 1957, in Gaskell, 2000). In terms of actually identifying the 'who, what, why and when' that a case study describes, documentary analysis can provide a rich and useful source of data. There is a large documentary corpus created by the project, both for formal and informal purposes, that provides a wide range of material for documentary analysis.

Documentary analysis (sometimes also 'qualitative document analysis') is a systematic procedure for reviewing and evaluating documents (Bowen, 2009) and can be used for a variety of purposes. Prior (2014) suggests that researchers can investigate both the content and use and function of documents. Content can be investigated both as a resource (content and thematic analysis) and as a topic (discourse analysis on how the document came into being). Use and function can similarly be investigated both ways via genre analysis and actor network theory.

Documents (which do not necessarily need to be, but often are, written texts) are non-reactive, stable data sources which have been established without researcher intervention (Bowen, 2009). They will not change with interrogation, and there is no Hawthorne Effect (although researchers are still susceptible to bias during analysis). They are created in naturalistic contexts (Bauer, Bicquelet and Suerdem, 2014). They are a practical, and in most cases, accessible, resource. Contemporary documents are widely available online, making their use cost-efficient. However, unlike a semi-structured interview or focus group, a researcher cannot steer the locus of the content. They may be (and in this case, are) extremely heterogenous. O'Leary (2014) categorises documents as those of public record (transcripts, reports, plans, manuals etc); personal documents (diaries, emails, social media posts etc) and physical evidence or artefacts, such as scripts, maps, posters, brochures etc. All of the documents here are of a 'public record' nature, although some of the material is meant for use within the consortium, some for use within the wider project, some for project reporting and other materials are created for a fully public audience.

Bowen (2009) recommends the consideration of issues derived from the discipline of History, such as the original purpose and target audience of the document, whether the author is writing of witnessed events or second-hand accounts, and the author's relationship to the material. O'Leary (2014) notes that 'latent' content must be observed as much as actual content. What is the tone? What is stated as fact? What is the agenda? A thorough comprehension of these underlies the credibility of research.

While document analysis is often used for triangulation or to generate questions or identify situations that merit further investigation, in certain cases, document analysis is the only or preferred method (Bowen, 2009). Such situations include when data that can no longer be

observed directly, or to track changes over time. Here document analysis is chosen for three reasons. It is used to capture the verity of a situation in which human participants might have incentives to characterise their actions or results in a way that does not accurately reflect the situation. For instance, in the background section it can be seen that City 1 defines open data as data, "*that is published (or will be published)*" on their website. In fact, whether data is published (and therefore available to everyone or not) is key in the definition. Further, as the research covers 3 years of the project, it removes the risk of unreliable or partial memory. Lastly, it covers a very wide range of topics, from project guidance, to piloting, to assessment, which might be difficult to cover in an interview. However, Bauer, Bicquelet and Suerdem (2014) warn that documents cannot be thought of as objective, as they are created by individuals with a specific purpose (which may include self-justification) and then selected and interpreted by the researcher.

On the other hand, documents are, unlike interviews, not constructed with the research agenda in mind. This means they will not provide all the necessary information, they may have gaps, or be inaccurate or inconsistent. Bowen (2009) notes that this may lead to having to search for more documents than initially accounted for. Other documents may not be accessible. Another criticism of the approach is that documents may not be an objective picture of reality, but may reflect the interpretations of the producers of documents (Denscombe, 1998). However, having sufficient documentation that relates not only to cities conceptualisation of their activities but also operational capture of their activity (i.e., presentations, deliverables and minutes) should help to allay this.

A key area to monitor is researcher bias. As a participant in the SCIFI consortium, I inevitably have previous views or experiences associated with each document. I therefore critically reflected on prior ideas about the content and meaning of each document to ensure I was not making unsupportable assumptions based on unreliable memory.

There are several generic approaches to document analysis. O'Leary (2014) has an eight-step process, which can be used with two different approaches, one where the researcher 'interviews' the documents, asking questions of them and finding answers in the text, and another using a quantified content analysis. Appleton and Cowley (1997) use a similar approach to the interview technique, devising 38 questions to be answered by each of the 77 documents under review as part of a five-step process of familiarisation of the data, simple sort, development of the criteria for critique, establishing a database and final analysis. Altheide and Schneider (1996) define a twelve-step process in 5 sections: documents, protocol development and data collection, data coding and organization, data analysis, and report. Bowen's (2009) four steps are finding, selecting, appraising and synthesizing documents, which are then subject to content analysis and/or thematic analysis. Evidently, while there is a certain amount of agreement on certain steps, there is also wide variation and room for interpretation. What is also evident is that the

authors agree that the actual analysis is only one step of a much longer process. I have adopted Appleton and Cowley's (1997) process, while using content and thematic analysis for analyzing the documents, as they are insufficiently similar for the critique and analysis tool developed by Appleton and Cowley (1997) to be used.

6.3.2 Document Data Collection

The SCIFI project begins with funding proposals in 2017 and has active documents from the project launch in mid-2018 to the launch of second civic accelerator in January of this year. The full corpus is constituted of hundreds of documents. As is common practice in these kinds of projects the documents were organised on the Google hosted shared drive by Work Packages (WP). There are also many hundreds of emails, and also websites owned by the cities that host their open data. The decision was taken not to include emails unless there was an issue of particular concern that required triangulation. This was partially because many would be about organisational matters, and partially because the body of emails would be limited to all emails from the project that I had access to - it was not pragmatic, or indeed, politic, to access emails in which I was not originally included. As it is neither practical nor sensible to review every document, it is vital to demonstrate a transparent selection strategy that avoids conscious or unconscious manipulation of the data.

This leads to the question of how many documents should be included? Sandelowski (1994) posits that the sample size must be small enough to manage the material and sufficiently large to provide new and rich understanding. Consequently, it is guided by the researcher's subjective judgment as the data is experienced and assessed, in relation to the research goals. Glaser (in Potts and Fugard, 2015) establishes theoretical saturation when no more new themes emerge. More concretely, Braun and Clarke (2006) suggest 10–100 for secondary sources (i.e. in this case, not interviews). This research covers 58 text documents and 35 spreadsheets or lists of metadata.

The first step in document analysis was to (re)familiarise myself with all the documents, which I did by reading through the entire shared drive. This led to an initial overview of which documents would and would not be included. There are 5 work packages:

Work Package 1 is a highly relevant work package, entitled 'Transnational Ecosystem for Open Data Innovation in the Public Sector'. This concerns the development of the selection of SMEs for the civic accelerator, via a challenge. It includes documents under 6 activities:

- A1.1 Map and assemble an innovation network
- A1.2 Local support groups
- A1.3 Identify, select and define challenges in the fields of mobility, energy and environment

- A1.4 Compile open data guidance based on top level harmonisation approach
- A1.5 State of the Art document on cross border market engagement
- A1.6 Draw on local support groups to gather city clusters per business case
- A1.7 Experiment with innovative public procurement competitions

Although an innovation network is a crucial element of open innovation it was decided that documents regarding how this was established were not a priority. The existence of the local support groups (other agencies and cities that are interested in the innovations) is important, but again, this only contains lists and records of communicating with them. The definition of the challenges is of great interest and individual documents here were reviewed for inclusion, as were the documents in A1.4. While issues of procurement are of great interest in open innovation in the public sector, as this is a limiting factor in many cases, procurement had not arisen sufficiently in the open data and open innovation literature to warrant its inclusion. The documents in A1.6 were mainly comprised of local language presentations (French and Dutch) to local cities, and so had to be excluded. Finally, although the name of the last activity was promising, it contained only a French language version of the contract between City 3 and the SME on the Watering Challenge.

Work Package 2 is also a key work package for selection of documents. Entitled Open Data Co-Innovation Acceleration Programme, it concerns the process of the civic accelerator. It includes documents under 8 activities:

- A2.1 Open up quality data needed for selected business cases
- A2.2 Co-develop solutions for public service challenges
- A2.3 Coaching and training in open public innovation, public private collaboration practices and living lab practices
- A2.4 Support and evaluation from city clusters
- A2.5 Transnational accelerator workshops
- A2.6 Establish accelerator method of operation
- Accelerator Programme 1 (All the documents for managing and delivering acceleration and monitoring the pilots in Call 1)
- Accelerator Programme 2 (All the documents for managing and delivering acceleration and monitoring the pilots in Call 2)

A2.2 contains a few documents about post-pilot procurement. A2.4 is empty, as is A2.5. A2.3 is planning for webinars and a survey for dissemination to SMEs on what training they would find

useful. A2.6 is the application form. Therefore, the key documents here are those pertaining to open data and the accelerator programme.

Only documents that focus on the pilot processes that were actually implemented have been included. There is no assessment of the selection process or analysis of the applications. This is for a number of reasons. The first is to reduce the amount of material. Approximately 130 organisations applied for the two Calls. For each SME applying this generated an application form, a presentation deck and a brief video, two sets of reviews of the application forms and a minimum of two sets of interview notes. The second is that contributions to the final decision were sometimes driven by other departments or agencies in the cities, and the relevant decision-making processes are difficult to access. The third is to reduce the role of the researcher in affecting the outcomes of the project. As a knowledge advisor to the project, I assisted City 3 in reviewing and interviewing the SMEs for the Watering Challenge in Call 1 and for the Parking and Waste Challenges in Call 2. Finally, while this corpus may make an interesting focus of study on its own merits, the literature review does not suggest key relevancy to the subject in hand.

Work Package 3 'Transnational Employment and Scale Up' is completely empty, as this part of the project does not commence until the early part of 2020.

Work Package 4 'Project Operations' contains the contract, reporting guidelines, budget and contact details, plus copies of the deliverables from the relevant WPs. The majority of these are either not relevant or contain personal data. The key documents in this work package are the Minutes of the fortnightly teleconferences, divided into 2 documents of 2017-2018 and 2019, and Minutes of the consortium meetings, which are held three times a year (8 at the time of research). Accompanying each set of consortium meeting minutes are the relevant presentations made by partners during the meetings. These cover a wide range of organisational, procedural and communications issues, so only presentations relevant to the development of the challenges, open data and the pilots were included.

I excluded all documents in Work Package 5 'Communications'. The majority of these documents were images for use in brochures, presentations and the website, agreed notes about messaging, communications and social media strategy documents, information about notifications to applicants or details of events that were being presented at. Content for the website was excluded because it all exists (apart from the images) elsewhere in documents. Where presentations were included, these were based on presentations I could find elsewhere in the drive. Below is a summary of the documents as they relate to activities across the project timeline.

Table 13 Summary of Documents in Relation to Project Phase

Date	Activity	Relevant Documents
July 2017	Project Launch	Data portals
January - June 2018	Development of business cases (challenges)	Business case outlines Open data policies and regulations
March 2018	Steering Committee 2	Minutes
June 2018	Steering Committee 3	Minutes
July 2018	Launch of digital innovation contest (Call 1)	Challenge web pages 2018
October 2018	Steering Committee 4	Minutes Metadatasets call 1 Contract Contract French version
January 2019	Launch of 1st civic accelerator (8 pilots)	FIWARE catalogue
January 2019	Steering Committee 5	Minutes Financing open data workshop
January - June 2019	Development of business cases (challenges)	Open data questionnaire Open data guidance package
April 2019	Steering Committee 6	Minutes Open data survey 1 Aims and value interviews Open data tracking Best practice and learning
July 2019	Launch of digital innovation contest (Call 2)	Challenge web pages 2019 Data inventories Roundtable report
July 2019	End of 1st civic accelerator	
October 2019	Steering Committee 7	Minutes Final pilot review presentations Metadatasets Call 2
January 2020	Launch of 2nd civic accelerator	Not included
January 2020	Steering Committee 8	Not included

6.3.3 Analytical Method

Appleton and Cowley (1997) note that before choosing the method of analysis of documents, it is, *"important for the researcher to carefully consider the purpose of this stage of the study"*.

The purposes, therefore, of this document analysis are:

1) To evaluate existing documents to describe their nature and content;

2) To assess how the cities are implementing open data for open innovation;

3) To understand how they are engaging with the open innovators around data;

4) To consider how what they do matches or deviates from the research (or their own stated intentions, where relevant).

Conversely, the purpose is not to assess 'are the cities doing open innovation well' or 'how innovative or successful are the outcomes' although these questions may be addressed as part of understanding whether some part of the open data provision or use affected this.

My first step was to familiarise myself with, *"the authorship, body and function of each document,"* (Appleton and Cowley, 1997). Post familiarisation with the documents, I conducted a 'simple sort' - dividing the documents up by their nature (Appleton and Cowley, 1997). This led to four categories - pilots, operations, open data and data lists (metadata). The documents in each category are defined in Tables 13, 14, 15 and 16 below.

6.3.3.1 Open Data

The following documents were identified as of possible relevance and interest.

Table 14 Open Data Related Project Documents

Document	Authorship, body and function
201904 OD Questionnaire [ODQ]	*Compiled results of the open data questionnaire devised by City 4 and distributed to cities in April 2019*
201904_ResultSurvey_OpenDataCity [ODQQ]	*Spreadsheet of charts of number of datasets opened and shared between cities and pilot SMEs, i.e., quantitative results from above, created by City 4*
Report - RoundtablesessionOpenData [SCRT]	*Notes by City 4 on 3 roundtables hosted on using open data for open innovation by the public sector at Smart Cities event, The Hague, 21.06.19. Content contributed by organisations outside the consortium. This has been included as (a). It was intended to gather external views in order to influence SCIFI's thinking, and (b). It was used to contribute to the last version of the Open Data Guidance Package deliverable.*
Tracking the open data publication in SCIFI [Track]	*Record of a teleconference hosted by City 4 to discuss the next steps for creating the second iteration of the Open Data Guidance Package deliverable 23.04.19*
WP1_A1.4_Workshop Open Data Guidance PartnerMeeting Cambridge [ODGSC05]	*Presentation at SC03 on the method to be used to prepare to open data. The goal of this guidance package is to enable city partners in the SCIFI consortium as well as all other cities to publish interoperable (open) data in such a way that the data can be used, reused and republished for value creation." Not used as normative.*
WP1-D1.4.1_Questionnaire-Open Data Guidance Package - Policies and Regulations [ODP&R]	*Answers to 15 questions on existing open data policies and regulations in each city, created at the beginning of the project (27.02.18) to inform the development of the Open Data Guidance Package deliverable by City 4. (These answers were used to inform the background of the cities presented earlier.)*

Document	Authorship, body and function
WP1_A1.4.1 Approach to enrich and improve ODGuidance [IODG]	*Planning document devised by City 4 to track the open data publication / release by cities and considering how to improve and enrich the open data guidance from January 2019 until the end of the project. Not used.*
Open Data Guidance Package v1 [ODV1]	*Project deliverable led by City 4 describing best practices for publishing open data and indicating areas the cities needed to address.*
Financing Open Data Workshop Bruges SC05 Collected Output [FODW]	*Comments and answers to 14 questions on financing open data portals presented in a consortium workshop. Transcribed and compiled from Post it notes created in the workshop. (UoS ran the workshop so did not input.)*

6.3.3.2 Pilot Process and Outcomes

The documents following pertain to the Challenge, Call and civic accelerator process and progress.

Table 15 Pilot Process Related Project Documents

Document	Authorship, body and function
[Phase - Pre-application]	
Call 1 Business Cases **Air Quality City 1** **Air Quality City 2** **Cycling – City 1** **Cycling – City 2** **Waste Management City 3** **Waste Management City 4** **Efficient Buildings** **Deicing** **City Centre Optimisation** **Watering**	*Completed business case template developed by each city for each challenge after consultation with stakeholders. There are more business cases than challenges that were eventually selected but only those that are selected for public challenges are included here.*
Call 1 - Smart Cities Framework Implementation Air Quality **Cycling** **Efficient Buildings** **Deicing** **Waste Management** **City Centre Optimisation** **Watering**	*Web pages developed to promote challenge. The format was agreed by the consortium, the content by the cities and UoS and the data listed by the cities. The challenges are promoted on www.smartcityinnovation.eu. They were created for the launch of each call (July 2018 and July 2019). The process of creation involved firstly the development of a business case for each open innovation, to which city officials and civil servants, academia, business and citizens contributed.* *The format of the page includes some generic background on the type of challenge (energy, mobility or environment); a description of the challenge as specifically understood by the city, expected outcomes, expected impacts and the datasets the city would make available for the pilot. This format was based on a similar format developed for the Horizon 2020 open innovation programme Data Pitch.*
Call 2 - Smart Cities Framework Implementation **Housing Transition** **Access and Parking** **Facilities Maintenance**	*Web pages developed to promote challenge. The format was agreed by the consortium, the content by the cities and UoS and the data listed by the cities. The challenges are promoted on www.smartcityinnovation.eu. They were created for the launch of each call (July 2018 and July 2019). The process of creation involved firstly the development of a business case for each open innovation, to which city officials and civil servants, academia, business and citizens contributed.*

Document	Authorship, body and function
Urban Logistics Vehicles **Multimodal Transport** **Bicycle Flows** **Shared Mobility Access** **Encourage Sustainable Commuting** **Pedestrian Flows**	*The format of the page includes some generic background on the type of challenge (energy, mobility or environment); a description of the challenge as specifically understood by the city, expected outcomes, expected impacts and the datasets the city would make available for the pilot. This format was based on a similar format developed for the Horizon 2020 open innovation programme Data Pitch. In 2019, a brief video of a city representative describing the challenge was added.*
[Phase - Application]	
SCIFI Agreement Including Annex 3 and 4	*Contract developed for use in procuring the pilots. There were some localisations to align with national and municipal law and language. This version is close to the original and is also in English.*
SCIFI Agreement Version Francais	*Contract developed for use in procuring the pilots. Localisations to align with national and municipal law and language, but also further clauses regarding intellectual property and sensor data.*
Aims	
[City 1] aims and value assessment **[City 2] aims and value assessment** **[City 3] aims and value assessment** **[City 4] aims and value assessment**	*Recorded conversations between UoS partner (not author) and various other consortium partners on how they identify and measure value in open data. These were created for the third iteration of the project deliverable 'Open Data Guidance Package' focused on value creation.*
[SME 1] aims and value assessment **[SME 2] aims and value assessment**	*Recorded conversations between UoS partner (not author) and various other consortium partners on how they identify and measure value in open data. These were created for the third iteration of the project deliverable 'Open Data Guidance Package' focused on value creation.*
Mid-pilot progress	
Best Practice and Learning April 2019 [BPLSC05]	*Preparation for workshop at SC05 for cities to share their progress and learnings so far from their pilots across the areas of procurement, data and pilot. The cities were asked to complete the template in advance. Three out of the four cities (excluding City 3) did so. Format was devised by UoS (not author).*
Progress Mapping 1 (notes on presentations in workshop at SC05 made by Soton) [MPPSC05]	*Notes from the above workshop to capture key ideas for use in later project deliverables. Made contemporaneously by author.*
End results	
SCIFI End Reporting Pilot Phase SME 4 [ERPP4] **SCIFI End Reporting Pilot Phase SME 6 [ERPP6]** **SCIFI End Reporting Pilot Phase SME 1 [ERPP1]**	*Templates completed by the SMEs to document what they feel has been achieved the last 6 months during the piloting of the solution. Not all available as the solutions involving AirQuality completed outside the accelerator due to sensor issues, consequently paperwork is delayed.*

Document	Authorship, body and function
SCIFI End Reporting Pilot Phase SME 3 [ERPP3]	
Pilot Presentation City 1 SC07 [PPC1SC07] **Pilot Presentation City 2 SCO7 [PPC2SCO07]** **Pilot Presentation City 3 SCO7 [PPC3SC07]** **Pilot Presentation City 4 SC07 [PPC3SC07]**	*Presentation by the cities on the end output of their pilots, learnings, and the next steps in terms of implementation. Originally presented at Steering Committee 07.*

6.3.3.3 Operations

The documents included under the heading 'operations' are minutes of the fortnightly teleconferences of the group and 6 of the 8 in-person consortium meetings. These documents cover all areas including both pilots and data, so it was decided to address these separately. These minutes, in Google Docs, are structured around the agenda of the two-day meetings, and are initially written by someone from the project management company, then reviewed by all members of the consortium. These documents, once finalised, are saved as PDFs, as they are official records of the project. Ultimately, given the sheer volume of material created by over 40 teleconferences a year, it was decided to only include the minutes from the partner meetings.

Table 16 Operations Related Project Documents

Document	Authorship, body and function
[SC01] Mechelen - October 2017	*No available minutes*
[SC02] Brussels - March 2018	*Written by project manager concurrently with meeting. Generally concerned with administrative harmonisation and mix of notation of comments and agreed actions. Approved by partners subsequently. Official minutes format.*
[SC03] Cambridge - June 2018	*Written by project manager concurrently with meeting. Focused on the agreement of and launch of challenges with associated data. Extensive discussion of the open data guidance deliverable, what best practice is comprised of and how cities can apply this in practice, rather than just theory. Approved by partners subsequently. Official minutes format.*
[SC04] Mechelen - October 2018	*Written by project manager concurrently with meeting. Extensive discussion about post-programme procurement and the interview process. Approved by partners subsequently. Official minutes format.*
[SC05] Brugge - January 2019	*Written by project manager concurrently with meeting. Extensive discussion about communications strategy and outreach. Part of the meeting was a workshop on open data that was captured separately in the open data documents. The pilots launched the day after this consortium meeting. Approved by partners subsequently. Official minutes format.*

Document	Authorship, body and function
[SC06] St Quentin - April 2019	*Written by project manager concurrently with meeting. Extensive discussion on development of challenges for call 2, changes that are required for the contract to be fit for purpose, and changes that need to be made to the accelerator process. Approved by partners subsequently. Official minutes format.*
[SC07] Mechelen - October 2019	*Written by project manager concurrently with meeting. Approved by partners subsequently. Still in draft format.*

Table 17 Metadata Related Project Documents

Document	Authorship, body and function
Challenges 2018 and 2019	*The part of this source important for analysis is the dataset list. The large part of the content was initially created by the city, but then rewritten for consistency by the University of Southampton partner. The datasets, however, were added using the city's own terms, phrasing and identification of ownership and openness.*
Metadatasets [META]	*Although I refer to all the data lists here as metadata, there are also specific metadatasets created as internal documents to the project. The purpose of the creation of these lists was twofold. The first was to meet the open data deliverable in the project. As this was anticipated to be a substantial amount of work, an interim deliverable of the metadata was agreed [Minutes, SC03]. The second purpose of the metadata is to provide information for the applicants as to what kind of data they may have access to, where data was not yet open, to describe what applicants might find in it [Minutes, SC03]. However, the metadata was not published openly in either Call. They are in Google Sheets. The template can be found in Appendix A.*
Data Inventories [DINV]	*These documents were intended to list all the datasets that were actually utilised or produced in the solution. The data inventories were developed retrospectively by the cities in Call 1, when it was realised that the solutions would include a variety of data types (open and closed), from a variety of owners (including proprietary datasets of the SMEs) and the types of solutions being piloted would create either data that might be personal or open data that might present a re-identification risk through triangulation. The purpose of creating these inventories was therefore for risk minimisation via review by a GDPR expert. In Call 1, the responsibility for creating the data inventory lay with the respective cities. However, one data inventory was provided to me by a SME working with City 4. At the time of writing, not all cities had finalised their data inventories. Webb (1984, in Appleton and Cowley, 1997) refers to such missing data as 'selective survival'. Two examples can be found in Appendix B.* *In Call 2, the decision was taken to transfer the responsibility for developing the data inventory to the SMEs from the cities, in order to increase timeliness and accuracy. This was made part of Milestone 1 in the pilots (and is therefore not yet available).*
Datasets Catalogued on FIWARE [FIW]	*FIWARE is a curated group of open source platform components which does not only include portal facilities but also other IoT components such as a context broker. The SCIFI Data Portal, based on FIWARE, was created to facilitate the implementation of solutions and to speed up the prototyping and experimentation process. FIWARE is an open alternative to existing proprietary Internet platforms and it is offered to SMEs to use it and get technical support when needed in the pilot phase of their solution. Some of the SMEs were quite expert in the use of FIWARE. Datasets were added to FIWARE through the process, and do not differentiate between Call 1 and Call 2. Currently City 3 has 11 datasets on FIWARE and City 4 has 8. City 1 has none and City 2 has catalogued its list of datasets as a dataset. There are no datasets generated in the pilots on FIWARE (although the intention was to publish them here).*
Data Portals [CDP]	*Datasets found listed on city portals. These are the key locations where the data that is published openly can be found for each city. Except for City 1, all cities published open data before the launch of the SCIFI project. A full list of the portals is found in Table 17.*

Table 18 List of City Portals and Associated License

City	Data Portal	Licence	Comment
City 2	https://www.city2.be/open-data	IPR limitation	
City 1	https://portaal-city1.opendata.arcgis.com/	License with each dataset	CC-BY
City 3	SCIFI FIWARE Portal http://opendata.agglo-city3.fr	License with each dataset	CC-BY
City 4	https://city4.dataplatform.nl	General terms of use on front page	

The metadata documents did not lend themselves to thematic analysis, but to a more pattern oriented, quantitative approach. Therefore, I separated the metadata documents from those pertaining to the pilots, operations and open data for analysis.

6.4 Case Study Data 2: Group Interview

As previously noted, documentary analysis has many benefits, but is strengthened by triangulation with other sources. The group interview has not been widely used, partly because of issues of access, and partly for the potential for bias and unwanted influence between participants. However, it is often used in economic contexts for product evaluation and for understanding work behaviour (Frey and Fontana, 1991). Such field interviews are useful for revealing perspectives and attitudes. This particular group interview is a field interview rather than a focus group, as the group is pre-defined (Frey and Fontana, 1991).

Participants in a group interview can 'bounce off' each other. This can result in obtaining more from the group session than from separate interviews. This also lends methodological rigour to the one on one process, by using participants for cross referenced multiple opinions (Frey and Fontana, 1991).

Not only does it support intersubjectivity, but vitally, it reduces the impact of the researcher's interpersonal relationship with the subject that arises in the one to one interview. This is important in this case as I represent the University of Southampton as knowledge advisor to the consortium. It is crucial that the interview avoided metamorphosing into a kind of 'test' where cities feel they are being judged on how well they have 'performed' open data. On the other hand, field interviews do require sensitivity to group dynamics. This was assisted by my in-depth knowledge of the group and time spent developing relationships over the past two years. Gaskell (2000) writes that the "*interview is a joint venture, a sharing and negotiation of realities",* which reflects the aim of its inclusion in my research.

Frey and Fontana (1991) note that both an informal, unstructured role or a formal, directive role can be assumed by the interviewer. I structured the group interview around certain findings of the document analysis. The aim of this was threefold: to confirm the accuracy of my findings; to obtain the views of city representatives on my findings, including whether the results reflected their understanding of reality, and to ask specific questions that had arisen in the document analysis, largely where motivations for divergence from rigid open data processes were unclear.

Further, the group interview format allowed me to gather the views of not only the cities but of the other participants in the consortium, the three other knowledge partners and the project manager. Their views on the results and process would be difficult to investigate during a one to one interview, as they might feel reluctant to comment on parts of the project that are not 'their' area. The final reason for using a group interview is for the practical reasons of limiting the time required during a consortium meeting to gather the data, and, once again, to limit the size of the corpus (Gaskell, 2000).

When using the group interview technique, the researcher needs to be aware that individuals in the group might feel stifled or conform, or there may be a high level of irrelevant data production. Bias can be increased if participants are influenced by one another's answers (Remenyi, 2011). However, being aware of bias is a crucial part of managing it. Below I have outlined the actual processes involved in the group interview data collection.

Table 19 Group Interview Process

Theoretical approach	Interview protocol
Pre-defined participants	The participants were 8 members of the SCIFI consortium representing 6 partners plus the project management company. The partners not represented were the University of Southampton and the Belgian networking partner, whose representative was new to the programme.
Physical format	The interviewees met in a room during a consortium meeting. The session lasted 1 hour. It was recorded.
Discussion topics	The topic guide for group interview was developed from the results of the document analysis and metadata analysis. The topic guide can be found below.
Analysis	The transcript was coded thematically.

Table 20 Topic Guide for Group Interview

Section	Content
Purpose	Confirmation of metadata analysis accuracy Views on metadata analysis Views on key issues that arose that diverge from open data practice
Timing	45 minutes – 1hour
Introduction	Welcome, explain format of session. Share participant information and consent forms.
Metadata Exhibits	Share metadata exhibits (ppt). Explain methodology. Gather responses.
Sensor Data	Share findings on sensor data (consent to use; access to; potential for being personal data). Gather views.

Personal Data	Share findings on personal data (sometimes exists within less sensitive datasets; potential for becoming personal data via use). Gather views.
Sharing Data	Share findings on data sharing (it seems easier than opening data; there is inconsistency over how this is managed in contracts). Gather views.
Closing Discussion	Is there anything I have not touched upon that you would like to add?

6.5 Thematic Analysis: Document Analysis and Group Interview

Following the establishment of the data as credible and authentic, the second step is to thematically analyse the data. Thematic analysis is a process for encoding qualitative information and can be used for all types of documents and interviews (Boyatzis, 1998). A straightforward model for thematic analysis is as follows; data familiarisation; initial code generation; theme search; theme review; theme definition and write up (Maguire and Delahunt, 2017).

The codes are outlined and described in a codebook (Appendix D). Each segment of data that is of interest or relevance is coded. There are two key approaches to thematic analysis (Braun and Clarke, 2006). The first is the 'top down' or 'theoretical' deductive approach. This is driven by the research question. The other is the 'bottom-up' data-driven inductive approach. However, a common approach is to use a combination of both, beginning with theory inspired categories and adding further inductive ones (Fereday and Muir-Cochrane, 2006). Accordingly, I began (Initial code generation) with top level codes from the framework created at the end of Chapter 4 – access, purpose, permissions, privacy and value - then developed further sub-codes inductively, by ascribing each theme that arose to a top level code. The general focus of each top level theme is shown in Table 20 below.

Table 21 Top Level Codes and Sub-theme Areas

Top Level Code	Sub-themes
Purpose	Regarding what data is used for, and who decides this; tracking reuse, or defining or limiting reuse in any way
Access	Regarding publishing, availability, discoverability or any issues that enable or restrict access to open data
Permission	Regarding licensing, ownership or other ways of granting permission, such as payment
Privacy	Regarding concerns about open data and personal data
Data User Value	Regarding how value is created, and what it looks like

Such a technique does not require line by line coding (Maguire and Delahunt, 2017). These codes were not predetermined, but were developed and modified as the process continued.

Some sub-themes arose, but lacked any informative value. (Not all themes fitted in a top level code: that were not related to a top level code were noted, but not included in the final analysis.) Analytical memos were used to support the iterative process (Maguire and Delahunt, 2017).

Theme definition was an important step in in the process. For instance, I had a number of pieces of text coded as 'Successful pilots' under the top level code of 'Value'. However, when I came to define the theme, it was no more substantial than a mention of a successful pilot. Consequently, it had little insight value, and was rejected.

The sub-themes (represented by the codes) presented an overview of activities, concerns, actions and views that fell within the main theme. Here the purpose was to understand what the range of activities, concerns, actions and views on each main theme was, as represented by the sub-themes, and what this illustrates in terms of compliance with, or deviation from, the archetype of open innovation with open data, as defined in the framework presented in Chapter 4.

This was a two step process – the first analytic coding was carried out only on the documents. A second analytic coding was carried out on the transcript of the group interview, which was informed by the results of the first coding, and also by the results of the content analysis.

Coding without a team is always at risk of subjective interpretation. However, given the non-changing format of the documents and the codebook, the research should be highly replicable, should a further researcher wish to assess the subjectivity. The codebook is available in Appendix D.

6.6 Content Analysis

Metadata is data about data. This category is documents that describe data used, or intended for use, in the pilots. They are essentially lists of datasets created for various purposes and at different times. It is hoped by including these in the analysis to triangulate what cities claim is open data with what actually is open data, as well as gaining further insights into ownership, availability and use. In the same way that the thematic analysis sought to explore whether the activities and views of the SCIFI project conformed to the idealised guidelines of the use open data, the content analysis of the datasets aimed to understand whether what was being used for open innovation was, as intended, open data published by the cities, or whether a wider spectrum of data was involved. This matters for a number of reasons, including understanding the barriers to the use of open data and also how organisations might be circumventing those barriers in practice.

Miles and Huberman (1994) state that the process for any qualitative research is collection, reduction, display, drawing and verifying conclusions. The primary form of display is the narrative.

However, analysing lists of datasets required a move away from this form, as used for the other categories of documents (Miles and Huberman, 1994). Ahmed (2010) suggests there are many other forms, and "*any way that moves the analysis forward is appropriate*". Accordingly, the data from the documents in this section are organised in tables, whose row headers are the codes developed from the close reading and content analysis.

A number of analytic approaches including thematic, content and text analysis were considered when deciding how to analyse the metadata. My research aim here was to understand if the labelling, categorisation or grouping of the datasets in each text was accurate, and from this to identify any further questions that this might raise.

The final analytic approach is derived from directed content analysis. The fundamental coding process in content analysis is to organize (generally large amounts) of text into considerably fewer categories. These categories are patterns or themes that are either directly expressed in the text or are derived from them analytically. Hsieh and Shannon (2005) state there are three forms of content analysis: conventional, directed, or summative. In directed content analysis, the study begins from theory (in this case, the definition of open data). Codes are defined prior to and during data analysis. Codes are derived from theory and relevant research findings (Hsieh and Shannon, 2005). Categories must be exhaustive and mutually exclusive. I established accuracy by creating categories, coding each dataset to a category, and then comparing this to the original labelling or classification of datasets. This comparison provides the theoretical relevance, and lifts the categorisation above purely descriptive information about the metadata (Haggarty, 1996)

The key difference, of course, is that I am not so much categorising "*units of communication*" (Haggarty, 1996) as re-categorising data on a number of dimensions to do with its publication. Although the categories are necessarily limited, I am deriving these from the types of data I identify across the metadata documents. It is a transparent process that can be easily replicated, and reliability and validity easily confirmed.

The documents included are Challenges 2018 and 2019; (16 web pages), Metadatasets (6 Google Sheets); Data Inventories (3 Google Sheets); Datasets catalogued on FIWARE (4 catalogue lists); Data Portals (4), and number of datasets opened and shared reported in Survey v1 (1) (a total of 35 documents).

Using the challenge datalists, metadatasets, SCIFI FIWARE portal and local government portals, I compiled and cross-referenced the data the cities wished to make available to the SMEs in the pilots, and their status in terms of both availability and ownership (whether they were open and under the control of the cities).

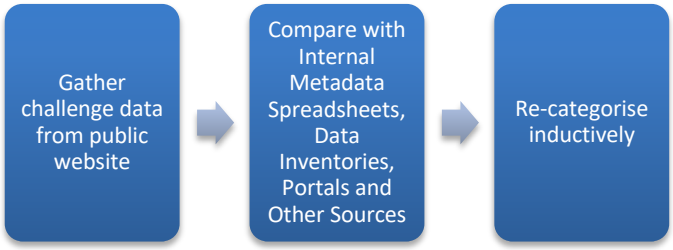

Figure 7 Overview of Categorisation Process for Metadata Analysis

6.6.1 Data Categories

The initial task was to devise categories of ownership and openness, and then to attribute the datasets correctly to each. The categories have been created from a content analysis of the various datasets and metadatasets, as described in section 5.4.1 in the previous chapter. The purpose of this analysis is to understand what data – in terms of the categories above - is being offered or identified for use by the cities (and to a lesser extent, whether it is categorised correctly by the cities). All the data provided for the pilots was initially intended to be open data belonging to the cities, in line with the mandate to open up datasets. Data that was eventually used in the pilots, held as Data Inventory lists, is discussed at the end of this section.From a review of the metadata categories were able to be derived as shown in the table below.

Table 22 Categories of Data Used in SCIFI Derived from the Project Metadata

Category	Sub-category
Ownership	City
	Third party
Availability	Open
	Closed
	Shared
Source	Existing
	Sensor
Content	Data
	Information
	Website

Each dataset is allocated a sub-category within each category; for instance, it might be city open sensor data, or third party closed existing data. Explanations of each of these sub-categories are below.

6.6.1.1 Ownership

City Data: The intent of the project and the cities was that the cities use their own open data to develop the services.

Third Party Data: In fact, not all the data eventually used (or even promoted on the Call website) is under control of the cities. In some cases (e.g., Safer Cycling) the data ownership is indicated on the website, but often not. These third parties are both private and public sector organisations, which own data.

6.6.1.2 Availability

Open: The term 'open data' appears to have had wide interpretations, and to have been aspirational rather than achieved. . The datasets in this category are currently published openly somewhere and linked to some form of license or open statement. However, this may have changed over the process of the call and pilot. At the point of the publication of the challenge the data might not have been freely published at this point, but instead be ready to be published. Alternatively, it might have been 'public' but not published openly.

Closed: This reflects several kinds of data. It might be data that the cities have not yet published, but it also may indicate data that the cities will never publish and will have to be shared. Within the project both this meaning and the status of data shifted over time so these meanings are all included in this section. In at least one case, a city described a third party closed dataset on the Challenge page.

The categories on the publicly available call websites have no standardised categories, although the intent is that these data sets are open either at the beginning of the call, or when the pilot commenced. In some cases, there is some ad hoc categorisation of 'not yet available' or links to portals or with owners appended, but this is inconsistent.

6.6.1.3 Source

Existing: The vast majority of the data is historical data which exists, in some form, somewhere, within the cities. (In one particular case, this was a pencil-written list of numbers in a desk drawer.)

Sensor: Two other kinds of data were listed. One was sensor data that was planned to be, or was collected, during the project.

Other: The third type was a manual count dataset that was intended to be created.

6.6.1.4 Content

Data: When analysing the metadata, it transpired what was classified as 'data' by the cities took a variety of forms. The vast majority of the datasets included in the various lists took the form of 'statistics collected together for analysis', but this was not the only type they promoted on the challenge websites. The datasets varied from three records to substantially more. No big data was included.

Information: Sometimes the data listed actually took the form of data visualisations or reports, which have been categorised separately.

This completes the section on Research Question 2. I now move on to discuss my approach to Research Question 3.

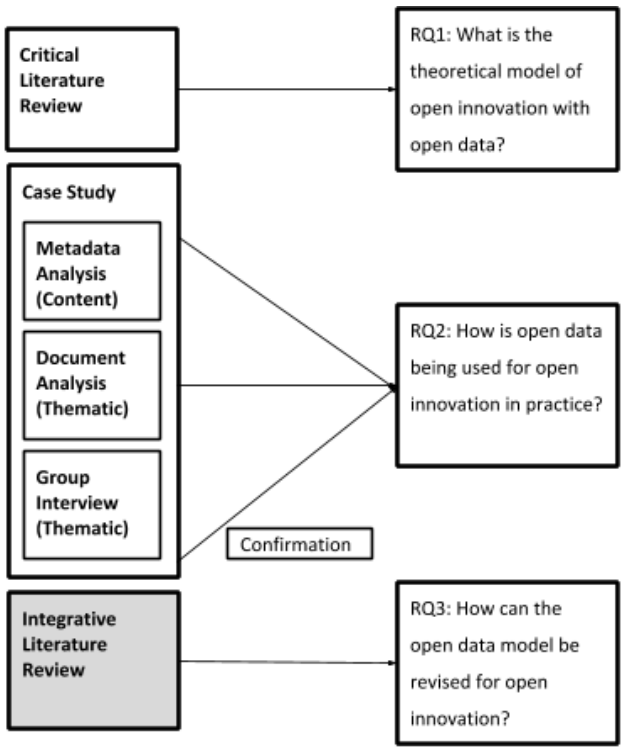

Figure 8 Outline of Research Methodology - 2

6.7 Research Question 3: Integrative Literature Review

Research Question 1 established a (simple) framework of open data and open innovation. Research question 2 examined this in action in a specific setting, and discovered substantial differences in how it was applied in practice, including data governance and privacy issues that led to the implementation of what Bacon and Goldacre (2020) refer to as, "*workarounds*". RQ3, therefore, attempts to develop a framework of data use in practice that reflects these issues. My theoretical orientation is that data sharing can inform compliant use of (open) data to offer an approach to innovation with data that could be followed by public sector innovators such as SCIFI. Therefore, I aim to review the literature on the features of open data that I have already

identified (Access, Permissions, Purpose, Privacy, Value) as they are found in the data sharing literature. Again, this is circumscribed as the data sharing literature that has an application for open innovation.

6.7.1 Defining the Integrative Literature Review

The method for RQ3 is an integrative literature review, "*a distinctive form of research that generates new knowledge about the topic reviewed,"* (Toccaro, 2005). A key purpose of an integrative literature review is the summarisation and comparison of terms or constructs about a particular phenomenon that create a base for theory development and concept definition (Webster and Watson, 2002; Whittemore and Kraft, 2005; Aburn, Gott and Hoare, 2015; Snyder, 2019). Although it can be used with mature topics it is also relevant for conceptualising and synthesizing emerging areas of study (Toccaro, 2005). It is useful for producing results that can have a direct application to practice and policy (Whittemore and Kraft, 2005).

While the systematic review is the most rigorous review method, this requires a narrow research question so is not an appropriate tool in this emerging area (Snyder, 2019). Further, the systematic nature means it is not appropriate for a diverse range of literature covering a range of methodologies and approaches (Whittemore and Kraft, 2005). The integrative review allows for varied perspectives and can embrace both theoretical and empirical studies (Whittemore and Kraft, 2005).

Threats to the validity of integrative reviews include scoping that is too broad, sampling that is ineffective and, in common with all qualitative research, inaccuracy or bias within the process (Jones-Devitt, Austen and Parkin, 2017). Research reviews such as integrative reviews are "*research of research,*" (Whittemore and Kraft, 2005). Consequently, they must meet the same standards of rigour of any other research method, including a specified purpose and variables of interest.

Toccaro (2005) emphasizes the importance of clarifying the need for a review of the literature. In this case, a framework has been developed from the decade-old open data literature. A departure from the framework has been identified in practice, which in particular, seems to reflect data sharing. Therefore, a review of the data sharing literature to understand how this might inform the framework is an apposite approach.

There are few guidelines for researchers performing integrative reviews, and successful reviews depend on skills such as critical and conceptual thinking. However, the basic choices for conducting a literature review still apply: designing the review, conducting the review, analysis and structuring and writing the review. There is no gold standard for evaluating the quality of papers in a literature review. Whittemore and Kraft (2005) suggest assessing the quality of papers where there is a discrepant finding, and discussing the quality in the final report.

6.7.2 Designing the Review

An initial scoping search was performed on an appropriate database (Aborn, Goff and Hoare, 2015). Following Watson and Webster (2002) I began with the Web of Science. An initial search was executed with the term "data sharing" in the topic. This returned 8,861 records. A further narrowing down to the title field returned 2,146 records. A review of the results showed that they were substantially composed of papers regarding the sharing of scientific data amongst the relevant communities (other scientists, patients), regarding the sharing and control of personal data by individuals (rather than collecting organisations) or otherwise pertaining to cross-border data harmonisation agreements. A further inspection showed that key known texts in the area were not being returned.

Subsequently the search was adjusted to focus on relevant sharing models for data-driven innovation in order to reduce the size and focus of the corpus. The types of data sharing used were *"data marketplaces"* (Richter and Slowinski, 2019) *"data commons"* (Fisher and Fortmann, 2010), *"data trusts"* (Hardinge and Wells, 2019) and *"data collaboratives"* (Susha, Janssen and Verhulst, 2017). The reasons for including these, and excluding others, is outlined in Table 29. It was desirable to have some commonalities in the types of data sharing, in that they should not be focused on one aspect (such as personal data, or the technical aspect) to the exclusion of all the others, to avoid, *"comparing apples with some fruit nobody's ever heard of"* (Keller, 2018).

Beginning with some experimentation on parameters, a search was performed in "Web of Science" for the terms, "data commons", "data marketplaces", "data trusts" and "data collaboratives" in the title. This returned too few results for 3 of the 4 terms. It was therefore expanded to topic. The time period was set at the default, 1970 - 2020. Whittemore and Kraft (2005) suggest that at least 2 to 3 search strategies should be used, as there are often inaccuracies in indexing. As well as Web of Science, a keyword based Google Search and citations were used (Jobin, Ienca and Vayena, 2019; Watson and Webster, 2002).

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram is used for depicting the selection of papers through the review. PRISMA is intended for systematic reviews (Moher et al, 2009). The full PRISMA statement is not appropriate for integrative reviews (Snyder, 2019). However, the flow diagram can still be successfully used for the literature selection (for example, Aborn, Gott and Hoare, 2015).

6.7.3 Conducting the Review

The PRISMA flow diagram differentiates between sources that are screened out and sources that are excluded for reasons of eligibility. Screening reviewed the title and abstract, to ensure that the source material focused on the correct area. Occasionally no abstract was available, in which

case the full text was reviewed where possible. Exclusions at this point included texts that primarily or largely addressed technical aspects of data sharing and texts in which the data sharing was subordinate to a major focus on medical or other scientific issues. Given the highly emergent nature of this area (especially around more recent models such as trusts and collaboratives) the decision was made to include some 'grey literature'. The decision as to what to include was based on Shopfel, (2011). This describes grey literature as, "*document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers*". This thus excluded blogs and newspaper articles. Two blogs in particular were frequently referenced in the literature: Scassa (2018) and McDonald and Porcaro (2015). However, to ensure the rigour of the process, these were not included. One paper was excluded for focusing on open data and included in the Chapter 2 literature review, one for focusing on personal data and a third for describing a workshop plan but no methodology or results. At the eligibility point, this excluded texts which were not available online (3) or via the University of Southampton logins (2) and which were not in English.

Table 23 Selected Terms for Search

Selected Terms	Justification of Inclusion
Data Marketplaces	Included in the Open Data Institute's Map of Data Access as core method (Keller, 2018)
Data Collaboratives	Focus of research initiative on social innovation and public-private data sharing innovation from the GovLab, NYU.
Data Trusts	Focus of research initiative on public-private/personal-public data sharing innovations from the Open Data Institute.
Data Commons	Included in the Open Data Institute's Map of Data Access as core method (Keller, 2018)

The full PRISMA flow is shown below.

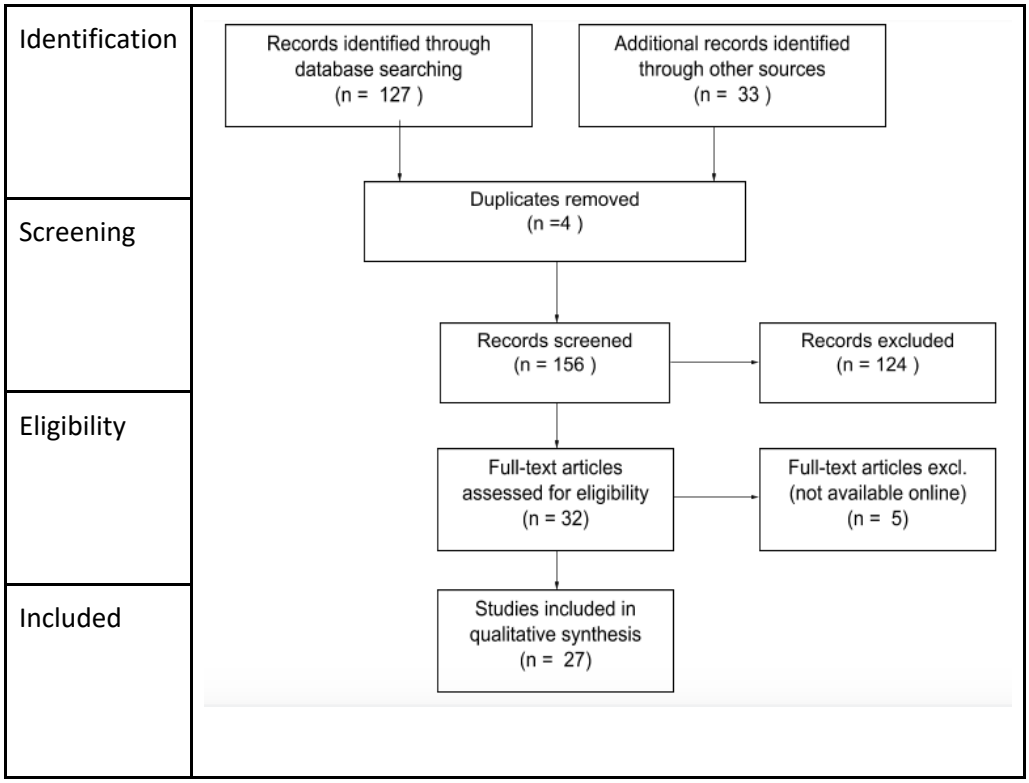

Figure 9 PRISMA Flow Diagram (Moher et al, 2009)

The included papers (after the first full paper review) are shown below.

Table 24 Papers Included in Integrative Review

Data Sharing Model	Paper or Report
Data Trusts (8)	Hardinges, J. and Wells, P., (2019). Data trusts will not be the final word on data sharing, but they might help. *Public Money & Management,* **39**(5), 320-321.
	Young, M., Rodriguez, L., Keller, E., Sun, F., Sa, B., Whittington, J. and Howe, B., (2018). Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing. In: *Proceedings of ACM (FAT'19)*. New York, NY: ACM.
	Stalla-Bourdillon, S., Wintour, A. and Carmichael, L., (2019). *Building Trust Through Data Foundations* [online]. Southampton: Web Science Institute. [21/01/20]. Available from: https://cdn.southampton.ac.uk/assets/imported/transforms/content-block/UsefulDownloads_Download/69C60B6AAC8C4404BB179EAFB71942C0/White%20Paper%202.pdf
	Hall, W. and Pesenti, J., (2017). *Growing the Artificial Intelligence Industry in the UK* [online]. London:Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. [21/01/20]. Available from: https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk
	O'Hara, K., (2019). *Data Trusts: Ethics, Architecture and Governance for Trustworthy Data Stewardship* [online]. Southampton: Web Science Institute. [21/01/20]. Available from: https://eprints.soton.ac.uk/428276/1/WSI_White_Paper_1.pdf
	Bunting, M. and Lansdell, S., (2019). *Designing decision making processes for data trusts: lessons from three pilots* [online]. London: Open Data Institute. [21/01/20]. Available from: http://theodi.org/wp-content/uploads/2019/04/General-decision-making-report-Apr-19.pdf
	Mulgan, G. and Straub, V., (2019). *The new ecosystem of trust: How data trusts, collaboratives and coops can help govern data for the maximum public benefit* [online]. London:Nesta. [21/02/20] Available from: https://www.nesta.org.uk/blog/new-ecosystem-trust/

Data Sharing Model	Paper or Report
	Stalla-Bourdillon, S., Thuermer, G., Walker, J., and Carmichael, L., (2020). Data Protection by Design: Building the foundations of trustworthy data sharing. *Data & Policy* **1**
Data Marketplaces (11)	Roman, D. and Gatti, S., (2016).Towards a Reference Architecture for Trusted Data Marketplaces: The Credit Scoring Perspective. In: *2nd International Conference on Open and Big Data, (OBD) 2016, Vienna, Austria, August 22-24, 2016.*
	Fricker, S. and Maksimov, Y., (2017). Pricing of Data Products in Data Marketplaces. In: Ojala A., Holmström, O. and Werder K. (eds) S*oftware Business. ICSOB 2017. Lecture Notes in Business Information Processing, vol 304.* Springer, Champagne, IL
	Stahl, F., Schomm, F., Vomfell, L., and Vossen, G., (2017). Marketplaces for Digital Data: Quo Vadis? *Computer and Information Science,* **10**, 22-37
	Stahl, F., Schomm, F., Vossen, G. and Vomfell, L., (2016). A Classification Framework for Data Marketplaces. *Vietnam J Comput Sci,* **3,** 137.
	Schomm, F., Stahl, F. and Vossen, G., (2013). Marketplaces for data: an initial survey. *SIGMOD Rec*. **42**(1), 15–26.
	Koutroumpis, P., Leiponen, A. and Thomas, L., (2017). The (Unfulfilled) Potential of Data Marketplaces, *ETLA Working Papers* 53 [online]. Helsinki: Etla. [21/01/20]. Available from: https://www.etla.fi/wp-content/uploads/ETLA-Working-Papers-53.pdf
	Carnelley, P., Schwenk, H., Cattaneo, G., Micheletti, G., and Osimo, D., (2016). Europe's Data Marketplaces - Current Status and Future Perspectives, European Data Market Study [online]. IDC: Luxembourg. [21/01/20]. Available from: http://datalandscape.eu/data-driven-stories/europe's-data-marketplaces-–-current-status-and-future-perspectives
	Richter, H. and Slowinski, P.R., (2019). The Data Sharing Economy: On the Emergence of New Intermediaries. *IIC,* **50**, 4–29.
	Duch-Brown, N., Martens, B. and Mueller-Langer, F., (2017). *The economics of ownership, access and trade in digital data; Digital Economy Working Paper* 2016-10 [online]. Seville:JRC Technical Report. [21/01/20]. Available from: https://ec.europa.eu/jrc/sites/jrcsh/files/jrc104756.pdf
Data Collaboratives (6)	Schwabe, G., (2019). The role of public agencies in blockchain consortia: Learning from the Cardossier. *Information Polity,* **24**(4), 437-451.
	Susha, I., Pardo, T., Janssen, M., Adler, N. and Verhulst, S., (2018). A Research Roadmap to Advance Data Collaboratives Practice as a Novel Research Direction. *International Journal of Electronic Government Research*, **14**(3), 1-11.
	Klievink, B., van der Voort, H. and Veeneman, W., (2018). Creating value through data collaboratives: Balancing innovation and control. *Information Polity,* **23,** 379–397.
	Susha, I., Janssen, M. and Verhulst, S., (2017a). Data collaboratives as "bazaars"?: A review of coordination problems and mechanisms to match demand for data with supply Transforming Government: people, process and policy, 11(1),157-172.
	Susha, I., Janssen, M. and Verhulst, S., (2017). Data Collaboratives as a New Frontier of Cross Sector Partnerships in the Age of Open Data: Taxonomy Development. In, *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2691–2700.
	Verhulst, S. and Sangokoya, D., (2014) . Mapping the Next Frontier of Open Data: Corporate Data Sharing. *Internet Monitor 2014: Data and Privacy*.
Data Commons (4)	Fisher, J. and Fortmann, L., (2010). Governing the data commons: Policy, practice, and the advancement of science. *Information & Management*, **47**(4), 237-245.
	Eschenfelder, K.R. and Johnson, A., (2014). Managing the Data Commons. J Assn Inf Sci Tec, 65, 1757-1774.

Data Sharing Model	Paper or Report
	Grossman, R., Heath, A., Murphy, M., Patterson, M. and Wells, W., (2016). A Case for Data Commons: Towards Data Science as a Service. *Comput Sci Eng.,* **18**(5), 10–20.
Other (1)	European Commission, (2018). *Towards a Common European Data Space* COM 232, Brussels.

The total number of primary sources for review was 27. The number of documents reviewed in the literature varies greatly, but this number or less is not uncommon (for example, Hopia, Latvala and Liimatainen, 2016, (10); de Souza, da Silva and de Carvalho, 2010 (5); Vagharseyyedin, 2016 (33)). Fricker and Maksimov (2017), reviewed in this study, used 11 out of an original 181 papers in their research.

6.7.4 Analysis

The goal of the analysis stage is an assiduous and balanced interpretation of the primary sources, coupled with novel synthesis of the evidence. I conducted a staged review, reading first the abstract then the full article (Torraco, 2005). The papers were read for their general argument but in particular assessed for any content that discussed how the data sharing model in question works. Although guided by the elements of the earlier open data framework - purpose, permission, access, privacy and value - I also sought to identify any additional concepts and understand their relevance. To assist with data reduction and data display, Webster and Watson (2002) suggest the use of a concept matrix, listing key topic concepts on one axis and the articles on the other, so relationships can easily be identified. This serves as a starting point for interpretation. The concept matrix is shown in Chapter 8. The constant comparison method for data analysis is suitable where varied data from diverse methods is used (Miles and Huberman, 1994). Data comparison enables the identification of themes, patterns, relationships, contrasts and evolving the particular from the general.

6.8 Summary

In order to investigate Research Question 2, How does the use of open data in open innovation in practice vary from the previously defined framework? I have analysed data collected from the Smart Cities Innovation Framework Implementation project.

Table 25 Description, Source and Contribution of Research Data

Data source	Description	Contribution
Project Dataset Metadata	All descriptions of datasets provided for use or used within the pilots developed in the civic accelerator	Inform understanding of how data is categorised with the project, and how this affects availability

Project Documents	Selection of documents created within the project for record keeping, information sharing and report making.	Generate structured insight into the relationship within the project between open data and its use, along 5 dimensions
Group interview	Views and thoughts of partners in SCIFI consortium	Reflective feedback from participants to triangulate previous results

Together, these sources represent substantial evidence for the wide variation in approaches to open data use. To investigate Research Question 3, How can comparison of other types of public and private data sharing arrangements inform the theoretical framework defined in RQ 1 so it more accurately reflects open data for open innovation as found in practice? I investigate selected types of data sharing structures. The aim is to understand to understand how their approach to access, purpose, permission, privacy and value can offer guidance on governance.

Chapter 7 Results - How does the use of open data in open innovation in practice vary from the previously defined framework?

This chapter presents the results for RQ2, 'How is open data being used for open innovation in practice?' The results are displayed in three parts, mapping on to the metadata lists, the pilot, operational and data documents and the group interview. The first part presents in tabular form the review and analysis of metadata documents from the project that compile or list data for the pilots. The purpose of this section is threefold: to allow clarity of review of contents of spreadsheets that would be difficult to comprehend in narrative form; to create new insights into the datasets that were used in the pilots; and to enable triangulation of insights about open data availability and use derived from the textual documents.

The second part presents the review and analysis of the remainder of the corpus in a more traditional narrative form. This focuses on five contributory elements of the original framework derived from the literature: the purpose of data use, access to data, permissions to use data, privacy and value.

The output of this group interview is presented as the third set of results, also in narrative form. The final results of the first part were presented to the participating cities for feedback during a workshop. The aim was firstly to clarify any points of opacity, and also to enable the cities to reflect on the 'reality' of how the data was used for open innovation, and to gain their commentary on this.

7.1 Results of Metadata Analysis

7.1.1 Call 1 Datasets

Call 1 was published in July 2018 and SME selection took place during October and November. At the time of writing, 7 SMEs had completed the civic accelerator. The tables below show the data as presented for use in the pilots on the Challenge website, and then the same data presented according to the categories of ownership, availability, source and content. This enables understanding of whether the data is truly open and available. The tables have been divided into Call 1 (2018) and Call 2 (2019) to allow for comparison and identification of changes or learning

Table 26 Datasets Identified for Use in Call 1 – BC1

BC1 - Safer cycling - Cities 1 and 2		
Datasets on Challenge page		City 2: Road Register: information on road infrastructure. City 2: Intersection Register: information on road intersections.

		City 2: Roadworks that affect traffic. City 2: Traffic measures: information on the amount of motorised and non-motorised traffic on specific roads. City 2: Demographic measures. City 2: Geolocation of popular citizen city destinations. City 2: Traffic accident information. City 1: Road infrastructure (street width, gradient, …) City 1: District borders City 1: Location popular destinations City 1: Number of school-age children City 1: Bike paths City 1: Cyclists by counting loops City 1: Location schools City 1: Number of accident reporting (police) City 1: Route2school (helps schools and municipalities by thoroughly analyse road safety on school routes and collect information about the travel behaviour of pupils) City 1: Bike counts
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	9	City 1: District borders City 1: Location popular destinations City 1: Number of school-age children City 1: Location schools City 1: Bike counts City 2: Traffic measures: information on the amount of motorised and non-motorised traffic on specific roads. City 2: Demographic measures. City 2: Geolocation of popular citizen city destinations City 2: Traffic accident information
Open Data – Third Party (Existing)	8	City 1: Number of accident reporting (police) City 1: Road infrastructure (street width, gradient, …) (Government) City 1: Cyclists by counting loops (Flowcontrol.be) City 1: Bike paths (Antwerp Province) City 1: Route2school (helps schools and municipalities by thoroughly analyse road safety on school routes and collect information about the travel behaviour of pupils) City 2: Roadworks that affect traffic (Flanders) City 2: Road Register: information on road infrastructure (Flanders). City 2: Intersection Register: information on road intersections (Flanders)
Closed Data – City (Existing)	0	
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	0	

All these datasets above, with the exception of the accident reporting, were simply listed on the challenge page as datasets, regardless of their ownership (which I have added). From the metadata for City 1 it is possible to see that they aimed to have data from more sources, including

closed such as Waze and shared such as Open Street Map, than they eventually ended up accessing or using (Appendix A).

City 2's metadata suggests they in fact aimed to use 9 datasets (of which two have more than one file), three of which are held by the Belgian geodata portal. Cities 1 and 2 both state in the open data survey [ODQ] that all the data sets were open and available before the pilots were started, although the metadata describes the Route2School data as being originally closed.

In the first Call, some cities shared the same challenge. However, although the Open Data Guidance Package [ODGSC05] recommended harmonization of the data sets, there was no compulsion to make all the datasets similar from both cities. The datasets provided by the two cities are quite different, with one focusing more on general traffic and road datasets, and the other on children, schools and cycling.

Table 27 Datasets Identified for Use in Call 1 – BC2

BC2 City 1 and 2 -Air Quality		
Datasets on Challenge page		City 2: Road Register: information on road infrastructure. City 2: Intersection Register: information on road intersections. City 2: Traffic measures: information on the amount of motorised and non-motorised traffic on specific roads. City 2: Particle measures: information from a mobile installation. City 2: Particle measures: information from a fixed installation. City 2: Demographic measures City 1: VMM, Flemish environmental society City 1: Road infrastructure (street width, gradient, …) City 1n: Public infrastructure inventory City 1: VITO MMM (independent Flemish research organisation in the area of cleantech and sustainable development) City 1: District borders City 1: Location of forests City 1: Curieuze neuzen (big citizen scientific research on air quality, specifically NO2) City 1n: Pollutant Release and Transfer Register (PRTR reporting)
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	7	City 1 Public infrastructure inventory City 1 District borders City 1 Location of forests City 2 Traffic measures: information on the amount of motorised and non-motorised traffic on specific roads. City 2 Particle measures: information from a mobile installation. City 2 Particle measures: information from a fixed installation. City 2 Demographic measures.
Open Data – Third Party (Existing)	7	City 2 Road Register: information on road infrastructure (Flanders). City 2 Intersection Register: information on road intersections (Flanders). City 1 Road infrastructure (street width, gradient, …) (Government) City 1 Pollutant Release and Transfer Register (PRTR reporting)) (Flanders) City 1 Curieuze neuzen (big citizen scientific research on air quality, specifically NO2)

		City 1 VMM, Flemish environmental society City 1 VITO MMM (independent Flemish research organisation in the area of cleantech and sustainable development)
Closed Data – City (Existing)	0	
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	0	

Again, all the data sets above were simply listed on the challenge page as datasets, regardless of their source. These are not differentiated by ownership, however, City 1's third party data has the name of the owning body attached in most cases.

Table 28 Datasets Identified for Use in Call 1 – BC3

BC3 - Waste Maintenance - City 3 and 4		
Datasets on Challenge page:		City 3 Location of approximately 700 cleanliness facilities Sectoral distribution of team in charge of cleanliness operation Location of green spaces Sensors to monitor prevalent conditions related to the maintenance of cleanliness facilities: percent of filling; problem with the bag distributor for dog poop collection Event calendar (cultural, sport events, etc.) User reports City 4 Events calendar Location of green spaces Location of bodies of water Street furniture Yet to be opened User reports Inspection reports on public spaces Garbage hotspots Number of people in city by date
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	5	City 3 Sectoral distribution of team in charge of cleanliness operation City 4 Events calendar City 4 Location of green spaces City 4 Location of bodies of water City 4 Street furniture
Open Data – Third Party (Existing)	0	
Closed Data – City (Existing)	7	City 3 - User Reports City 3 Location of approximately 700 cleanliness facilities (some restricted attributes) City 3 Location of green spaces (some restricted attributes)

		City 3 - Event calendar (cultural, sport events, etc.) (some restricted attributes) City 4 - User reports City 4 - Inspection reports on public spaces City 4 - Garbage hotspots
Closed Data - Third Party (Existing)	1	City 4 - Number of people in city by date
City Sensor Data	1	City 3 - Sensors to monitor prevalent conditions related to the maintenance of cleanliness facilities: percent of filling; problem with the bag distributor for dog poop collection
Third Party Information	0	

City 4 differentiates between data that is already open (albeit not giving a great amount of detail) and data that is yet to be opened. However, it did not distinguish publicly between data it held, and data held by another body (in this case, a Stichting, or foundation). City 3 does not publicly distinguish between the open or closed nature of the datasets, but it can be seen from the metadata that User reports (Citizen reports) and the localities and certain properties of the green spaces are not open and require some internal discussion. Further, City 3 did not have an open data portal at the beginning of the project, so it is questionable whether the 'Sectoral distribution' was fully open. (City 3 did not select anyone for the pilot so these were eventually not utilised). Equally, the location of the cleanliness facilities (essentially wastebins) also has some restricted attributes. Again, City 3 does not differentiate in the public domain between existing data and data that may be created during the pilot (sensor data).

Table 29 Datasets Identified for Use in Call 1 – BC4

BC4 - City 1 - Building Efficiency		
Datasets on Challenge page:		District borders Ownership, age of building (kadaster) Electricity & gas use (Eandis) Solar panels (Eandis) Thermographic map Flemish solarmap
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	1	District borders
Open Data – Third Party (Existing)	4	Electricity & gas use (Eandis) Solar panels (Eandis) West Flemish thermographic map Flemish solarmap (www.vito.be)
Closed Data – City (Existing)	0	
Closed Data - Third Party (Existing)	1	Ownership, age of building (kadaster) (restricted)

City Sensor Data	0	
Third Party Information	0	

The challenge largely depended on open data from other sources that was gathered by the city, one dataset of which was closed. No SME was selected for this pilot, so no data was opened.

Table 30 Datasets Identified for Use in Call 1 – BC5

BC5 - Watering Optimisation - City 3		
Datasets on Challenge page:		Events calendar related to occupation of public green spaces; Location of green spaces; Weather datasets from weather station owned by of the local government or from collaborative weather network if needed to gather forecast weather; Prevailing environmental conditions of public green spaces through sensors (air humidity; ground humidity)
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	0	
Open Data – Third Party (Existing)	1	Weather datasets from weather station owned by the local government or from collaborative weather network if needed to gather forecast weather;
Closed Data – City (Existing)	2	Events calendar related to occupation of public green spaces;(some restricted attributes) Location of green spaces; (some restricted attributes)
Closed Data - Third Party (Existing)	0	
City Sensor Data	1	Prevailing environmental conditions of public green spaces through sensors (air humidity; ground humidity) Sensor data will be collected during pilot
Third Party Information	0	

This challenge in fact presents no city open data. On the challenge page these datasets are presented without comment except for the note that the sensor data will be collected during the pilot. From the metadata, it can be seen that some attributes of the calendar will have to be restricted before publication. Along with the waste management pilot, the list explicitly includes data that will have to be collected during the pilot, which does not exist at the point of publication of the challenge. Post development of the solution all datasets are apparently available with the CC-BY (on the FIWARE portal) although some attributes are restricted. It is not clear whether these have been removed from the online version, in which case the data usefulness has been degraded, or it is there and not for public consumption, in which case it has been erroneously listed with an attribution license. For instance, according to the Open Data Survey results, City 3 claims one data set (the calendar of occupation) continues to be shared, while it is also available on the FIWARE portal. Associated documentation pertains to the API, rather than the data content. This will be checked in the group interview.

Table 31 Datasets Identified for Use in Call 1 – BC6

BC6 -City Centre Optimisation - City 4		
Datasets on Challenge page:		Tree locations Green locations Bodies of water Street furniture National walking routes Regional biking network Monuments Liveability neighbourhoods Noise Not yet opened Drinking water fountain locations Outdoor fitness equipment locations Playground locations Public sport grounds locations Priority cycle lanes
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	7	Tree locations Green locations Bodies of water Street furniture Monuments Liveability neighbourhoods Noise
Open Data – Third Party (Existing)	2	National walking routes Regional biking network
Closed Data – City (Existing)	5	Drinking water fountain locations Outdoor fitness equipment locations Playground locations Public sport grounds locations Priority cycle lanes
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	0	

City 4 differentiated on the challenge page on the basis of availability, but not on the basis of ownership (the priority cycle lanes may also actually be a regional or national dataset, as it is not found in the metadata list this is quite likely). It is not clear who owns the 'noise' and 'liveability neighbourhoods' data, as these were not included in the metadata either. As above, as this challenge did not have a pilot, it is harder to locate the relevant information about the data.

Table 32 Datasets Identified for Use in Call 1 – BC7

BC7 - De-Icing Optimisation - City 4		
Datasets on Challenge page:		Salt spreading routes Major road traffic densities Traffic accidents Cycling accidents Regional cycle network Heightmap Local road traffic densities Priority cycle lanes
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	4	Salt spreading routes Major road traffic densities Cycling accidents Local road traffic densities
Open Data – Third Party (Existing)	3	Regional cycle network Heightmap Priority cycle lanes
Closed Data – City (Existing)	1	Traffic accidents
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	0	

For this challenge, City 4 publicly states that all the data sets are available. However, a search on the data portal for 'verkeersongeluk' (traffic accident) and 'fietsongeval' (bicycle accident) returned no results. The accident datasets are not listed in the metadatasets. There is no reason stated for this, so it may be that they could not in fact be opened. Similarly, there is no record of priority cycle lanes, however, it seems likely this is again because this data set does not belong to City 4.

Table 33 Number of Datasets Reported as Opened and Shared in Call 1

City	Datasets published as open data	Datasets shared	# of pilots
City 2	15	0	2
City 1	16	0	2
City 3	8	1	1
City 4	23	4	2
TOTAL	62	5	7

These are compilations of the answers to the questions in an internal Open Data Survey conducted by City 3 during the first pilot in April 2019. The questions were 'What is the number of datasets published as open data (in SCIFI) so far?' and 'What is the number of (still closed) datasets shared directly with the solution provider?'

7.1.2 Call 2 Datasets

Call 2 was published in July 2019 and SME selection took place during October and November. At the time of writing, the cities had identified SMEs they wished to work with, but the pilots had not yet begun. This means that data that will eventually end up as open may not be available on portals currently.

Table 34 Datasets Identified for Use in Call 2 – C1

C1 - Encourage Sustainable Commuting - City 1		
Datasets on Challenge page:		Road Congestion GPS Tracking Global Traffic Scorecard
ANALYSIS		
Datasets:	#	Description
Open Data – City (Existing)	0	
Open Data – Third Party (Existing)	0	
Closed Data – City (Existing)	0	
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	3	https://ec.europa.eu/transport/facts-fundings/scoreboard/compare/energy-union-innovation/road-congestion_en https://www.tomtom.com/en_gb/traffic-index/ranking/ http://inrix.com/scorecard/

City 1 did not present any datasets for use in this challenge. The information was visual rather than raw data, sourced from Tom Tom and the Joint Research Centre; a Tom Tom index rather and an Inrix (data and insight provider) index. There were 5 interviews for this challenge; none of the applicants shortlisted by the city representatives in the project were approved by the relevant agency in the city. No metadata list was available for comparison. In Call 2, some datasets were linked from the page so that applicants could explore them further.

Table 35 Datasets Identified for Use in Call 2 – C4

C4 – Shared Mobility Access - City 1		
Datasets on Challenge page:		Key figures from the strategic advisory board for the Mobility and Public Works policy area General figures about the city and its citizens Article about the need for data to support the rights of older persons with disabilities Website to inform elderly people about mobility, living, social relations, health, income Facts and figures from the advisory and participation body for the elderly at the Flemish government
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	0	
Open Data – Third Party (Existing)	0	
Closed Data – City (Existing)	0	
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
City Information	1	General figures about the city and its citizens
Third Party Information	3	https://www.mobiliteitsraad.be/mora/thema/kerncijfers Facts and figures from the advisory and participation body for the elderly at the Flemish government (404) Key figures from the strategic advisory board for the Mobility and Public Works policy area

The city did not differentiate between websites and data, and did not present any datasets for this challenge. It provided direct links to the information and websites. There were no applications suitable in shared mobility.

Table 36 Datasets Identified for Use in Call 2 – C9

C9 – Urban Vehicle Logistics - City 1		
Datasets on Challenge page:		Datasets available soon Location of garbage cans Planning input Vehicle descriptions (capacity, options, …) Historic garbage can emptying statistics
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	0	
Open Data – Third Party (Existing)	2	Road Register: information on road infrastructure. Roadworks affecting traffic (Flanders, XML)
Closed Data – City (Existing)	4	Location of garbage cans Planning input

		Vehicle descriptions (capacity, options, …) Historic garbage can emptying statistics
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	0	

While the city does not differentiate between ownership, it provides the two open datasets with links to the Flanders Geoservices website. (These were broken when they were accessed after the call was closed.) The four datasets belonging to the city are noted as being 'available soon'.

Table 37 Datasets Identified for Use in Call 2 – C5

C5 – Multimodal Transport - City 2		
Datasets on Challenge page:		7 km club Bike to work Fietsersbond Demographic Measures Geolocation of popular citizen city destinations Road Register: information on road infrastructure Traffic measures: information on the amount of motorised and non-motorised traffic on specific roads
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	3	Traffic measures: information on the amount of motorised and non-motorised traffic on specific roads Demographic Measures Geolocation of popular citizen city destinations
Open Data – Third Party (Existing)	1	Road Register: information on road infrastructure
Closed Data – City (Existing)	0	
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	2	Fietsersbond (website) Bike2Work Club (website)

Again, City 2 reduced the number of datasets presented. Although some of the Call 1 datasets might have been appropriate for this challenge, they are not presented here and it is not possible to assume that potential applicants would find them in the previous call information. No metadatasets are available for cross referencing. City 2 appointed 2 SMEs with which to co-create pilots.

Table 38 Datasets Identified for Use in Call 2 – C7

C7 – Access and Parking - City 3		
Datasets on Challenge page:		Weather dataset Pricing policy: streets with paid parking and tariff Parking spaces reserved for disabled people Event calendar issued Tourism Office Datasets available soon ID number and location of time stamp Number of parking slot covered by the time stamp Parking spaces reserved for disabled people Event calendar issued Tourism Office
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	3	Pricing policy: streets with paid parking and tariff Parking spaces reserved for disabled people Event calendar issued Tourism Office
Open Data – Third Party (Existing)	1	Weather dataset
Closed Data – City (Existing)	2	ID number and location of time stamp Number of parking slot covered by the time stamp
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	0	

The city data not publicly available here is listed under 'Datasets available soon'. The ID location and time stamp may, if combined with information such as parking spaces for disabled people and other items on the map, start to create personal data, so this may never be opened, dependent on what is done with it. No potential sensor data is listed here, unlike in Call 1, even though the parking solutions are likely to use this.

Table 39 Datasets Identified for Use in Call 2 – C3

C3 – Pedestrian Flow - City 3		
Datasets on Challenge page:		Location of main points of interest in the city Weather dataset Event calendars issued Tourism Office Streets names and locations
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	3	Location of main points of interest in the city Event calendars issued Tourism Office Streets names and locations

Open Data – Third Party (Existing)	1	Weather dataset
Closed Data – City (Existing)	0	
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	0	

The city published 4 of the datasets in the challenge, however, another events dataset was in the metadata list as open but not publicly listed.

Table 40 Datasets Identified for Use in Call 2 – C8

C8 – Waste collection - City 3		
Datasets on Challenge page:		Weather dataset Localisation of public cleanliness facilities including ashtrays Inventory and localization of green spaces Event calendar issued Tourism Office Citizens report
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	3	Localisation of public cleanliness facilities including ashtrays Inventory and localization of green spaces Event calendar issued Tourism Office
Open Data – Third Party (Existing)	1	Weather dataset
Closed Data – City (Existing)	1	Citizens report
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	0	

Only two of the datasets mentioned (inventory of waste bins and inventory of green spaces) are mentioned in the metadata set. However, all except the citizens' reports are openly available on the project dataportal. All datasets were listed as open.

Table 41 Datasets Identified for Use in Call 2 – C6

C6 – Bicycle Flows - City 4		
Datasets on Challenge page:		Green Areas BGT the basic geographic datasets of the city consisting of buildings, roads, etc

ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	4	Green Areas Tourist Biking Routes Traffic intensity Main tourist walking route
Open Data – Third Party (Existing)	4	BGT the basic geographic datasets of the city consisting of buildings, roads, etc. (NL) PDOK is the European open government geodata platform Cycling routes national Student numbers (City Technical University) Comfortable Cycling Roads (National Government)
Closed Data – City (Existing)	1	Location of bicycle traffic lights
Closed Data - Third Party (Existing)	3	Bike rent - Mobikes, NS Routes of bikes rented - Mobikes, NS Number of bikes collected by regional bike depot because wrongly parked (Haagland depot)
City Sensor Data	0	
City Other Data to be generated	1	Number of cyclists (manual counting)
Third Party Information	1	Mobility data model Metropole Rotterdam Den Haag

The city only promoted two datasets on the challenge page. However, there were 14 listed in the internally distributed metadata list. While not all of this was either available or within the control of the city, there were substantially more open datasets than advertised. Consequently, this diverged from the actual situation, but unusually, by under promoting the available data.

Table 42 Datasets Identified for Use in Call 2 – C2

C2 – Housing Transition - City 4		
Datasets on Challenge page:		BAG, basic registration addresses and (use of) buildings Demographic data (neighbourhoods, age, education, etc. per area and for the whole city) Basic registration topography WOZ – Cadastral Value of properties Datasets available soon Building permits Building projects overview
ANALYSIS		
Datasets:	**#**	**Description**
Open Data – City (Existing)	1	Demographic data (neighbourhoods, age, education, etc. per area and for the whole city)
Open Data – Third Party (Existing)	3	BAG, basic registration addresses and (use of) buildings (Dutch Cadastral) Basic registration topography (Dutch Cadastral) WOZ – Cadastral value of properties (Dutch Cadastral)

Closed Data – City (Existing)	2	Building permits Building projects overview
Closed Data - Third Party (Existing)	0	
City Sensor Data	0	
Third Party Information	0	

In the metadata lists only the demographic data and the cadastral data are listed. The city does not differentiate between their data and third parties, but does note which datasets are not currently available (i.e., correctly reflects the situation).

Table 43 Total Number of Datasets Identified for Use by Category and Call

Data type	Call 1	Call 2
CITY OPEN DATA	34	17
THIRD PARTY OPEN DATA	24	13
CITY CLOSED DATA	13	11
THIRD PARTY CLOSED DATA	2	3
CITY SENSOR DATA	2	0
THIRD PARTY INFORMATION	0	9
CITY INFORMATION	0	1
CITY OTHER	0	1
TOTAL	75	48

According to the Open Data Survey carried out half way through the first civic accelerator, despite the wide range of ownership and availabilities, the cities provided a total of 62 datasets to the SMEs, with 5 being shared.

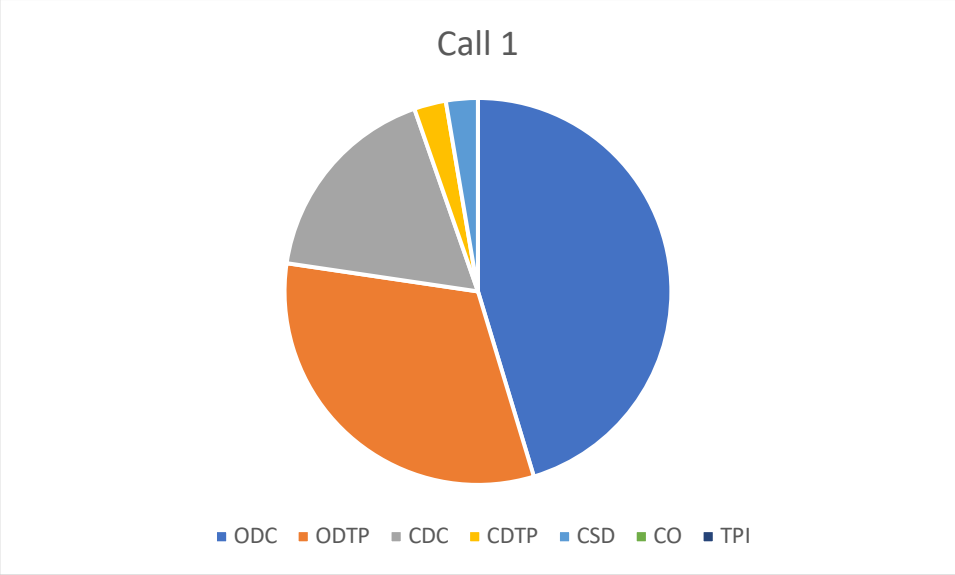

Figure 10 Data Identified for Use in Call 1, By Type

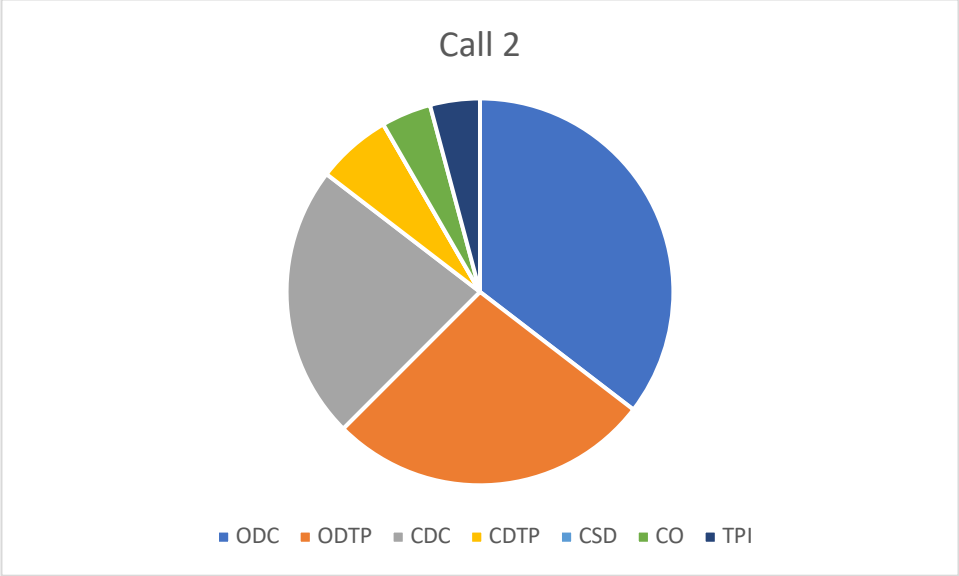

Figure 11 Data Identified for Use in Call 2, By Type

Overall this gives a clear picture of a wide variety of data being intended for use, not just open data belonging to the cities. In the second call, more types are included than in the first call. Although not included in any of the data lists, other documents [SME2] show that at least one pilot also used proprietary data.

7.2 Key Issues in the Metadata Analysis

The quantitative findings above are now explored in more detail.

7.2.1 Categories of Availability

There are more categories of both ownership and openness than acknowledged publicly on the SCIFI website, as is shown by the analysis above. For example, in Challenge 1 only one city differentiated between open and non-open data sets. Cities are aware of the difference between publicly available data and open data - in the Open Data Guidance v1 Appendix 4 [ODV1] 14 questions are provided to assist cities in establishing whether the datasets are public and whether they can be published as open data - clearly separating the two steps. Despite this when completing the metadata template some cities have filled this in as '*public*' and some as '*open data*' [META].

Overall it can be seen that only around half the data is actually owned by the cities and open for the beginning of the digital innovation contests and even the pilots, and the challenge pages do not accurately reflect the situation.

City 3 utilised a dataset that was 'public but restricted in some attributes' (and ended up being shared). In Call 2, City 1's 'datasets' are, on the whole, not data, and none are city open data.

7.2.2 Categories of Ownership

The cities do not differentiate between their data and that of third parties in the promotion of the challenges on the project website, or consistently attribute this. There is no pattern in deciding when they acknowledge the data belongs to a third party and when they do not - for instance, City 1 clarifies that some of the data in the Air Quality challenge belongs to parties such as VMM (the Flemish Environment Agency), and that the traffic incident data belongs to the police in the Cycling challenge, but not that the bike path data in the Cycling challenge belongs to Antwerp Province.

In Call 2 it does not mention the owners of the various indices it links to. This imprecision is also reflected in other documents: City 2 say they have opened 'about 15' data sets when replying to the April 2019 Open Data Survey [ODQ]. City 1 specified a dataset owned by the national government, and which was closed. In this case, it was not needed as the pilot was never run.

7.2.3 Comparison with Data Inventories Used in Pilot Solutions

Of all the data offers above, only three can be compared with the data that was used in the piloted solution. These are for challenges BC 3, 5, 7 in the first call.

The pilot of BC5 claims to use 5 datasets (4 were suggested). These are the weekly calendar of sports activities on the fields; two sensor measurements (humidity and ambient temperature), the weather forecast and a battery sensor measurement. It seems likely, however, that a further two GIS datasets were included, to identify where the fields were. Of these datasets, the weather is published as open data by third party provider, and events calendar appears published as open data by the city. However, the calendar is restricted for technical and security purposes. The inventory states, "*Not possible to publish this dataset because of security concerns according IT department. So they remain closed data,*" [BC3 DINV]. There is therefore some confusion surrounding this. The three sensor datasets created by the pilot remain closed as they are experimental. However, the sensor datasets are published to the SCIFI hub, but not openly. This means that the results of the open data survey, which reflects what was initially shared, and states only one data set remains closed, are not aligned with the final inventory.

The data inventory of BC3 for City 4 shows that 22 datasets were considered or requested for use (again this does not include a basic map, which it must be assumed is used). Eight datasets were initially presented for consideration in the challenge. Of the 22 requested, 6 datasets were not provided. This was mainly because either the city themselves were not able to provide them (such as student housing maps) or because they were sensor datasets which were not yet created. Of the datasets that were requested and delivered, there is a variety of statuses, including data that is not open but was shared and data that was open but not published anywhere and so was

shared. In all, there were 4 datasets that were characterised as open, but were not published, so were (at least initially) shared.

SCIFI Datasets for Waste Management in City 4					
Maintenance					
Name applicant	Dataset (ENG)	Owner	Open data	Published	Published where (URL of
SME1	Locations of public litterbins	City 4	Y	Y	City 4 data portal
SME1	Network of roads in the city	Rijkswatersta	Y	Shared	
SME1	Location of green spaces	City 4	Y	Publication	City 4 data portal
SME1	Events calendar + Locations	City 4	Y	Shared	City 4 website
SME1	Demographic data	City 4	Y	Y	https://www.cbs.nl/en-
SME1	Intended use for buildings	City 4	Y	N, shared	Basisregistratie
SME1	Building types in the city	City 4	Y	N, shared	Basisregistratie
SME1	Locations of markets in the city	City 4	Y	N	
SME1	Type of urbanized area	City 4	Y	Y	http://data-
SME1	Tourist attractions	City 4	Y	N, shared	
SME1	Map of elevation	City 4	Y	Y	PDOK
SME1	Citizen reports on public litterbins or	City 4	N	Shared	
SME1	Distribution of the team for collecting	Werkse!	N	Shared	
SME1	Number of people in city by date	Stichting	N	N	
SME1	Ages of the buildings in the city	City 4	N	N, shared	Basisregistratie
SME1	Commercial activities	ntb	N	N, shared	
SME1	Benches		Y	N, shared	
SME1	Picknick tables		Y	N, shared	
SME1	Bus stops		Y	N,shared	
SME1	Play grounds		Y	Y	

Figure 12 SME Data Inventory for Waste Challenge

The inventory provided by the SME in this pilot does not match that held by the city. It lists 20 datasets, of which 16 are held by the city and 4 belong to other bodies. While it suggests that 4 non-open datasets where shared, it also suggests that some datasets characterised as open were 'shared' rather than published. It also notes that a dataset that was originally promoted on the challenge web page was unable to be opened or shared, as it is held by a third party.

For BC7 16 datasets were listed in the city data inventory. Of these, six were not utilised in the project (so should not have been included). These included traffic accident reports (owned by the police); salt-spreader sensor information (owned by the company who runs the salt-spreaders) and CCTV, which clearly carries personal data risk. Four datasets were shared while work went on to define whether they could be opened. This is confirmed in the open data survey results: *"4 datasets were shared because these datasets were not collected and stored in a proper way. We are now working on that before opening up. Thus, the start-ups have to work from a single datadump."* [ODQ]

7.2.4 Amount of Data Made Available

In Call 2 there appears to be a distinct reduction in the number of datasets promoted. This is despite a comparable amount of challenges in both rounds (9 with no shared challenges in Call 2, 7 with 3 shared challenges in Call 1). City 4's metadata showed it had considerably more datasets in mind than they promoted on the website. During the publication of the call (July 1 - October 1)

cities were repeatedly reminded that they could update the datasets on the website (Minutes, July, August 2019). In general, in both City Open Data and third party open data, in Challenge 2 the cities promoted only half the amount of data.

This is particularly striking for City 1, who promoted no open data at all in Call 2. City 2 also hugely reduced the effort put into presenting datasets, even where some of the Call 1 datasets might have been appropriate. In Call 2 there were 9 challenges but none were shared.

7.2.5 Data in Shared Challenges

The Open Data Guidance Package v1 created by City 4 for the project recommends, *"If cities work together in one challenge, it is recommended they analyse the datasets with similar data and strive to harmonize the datasets as much as possible for solution development,"* [ODV1]. In practice, this was achieved very well in 1 out of the 3 shared challenges, reasonably well in another and not at all in the third.

In the Waste challenge in Call 1, Cities 3 and 4, despite being in different countries, managed to replicate their datasets quite well. However, only City 4 selected a SME to work with, as City 3 did not find an approach they liked.

In the Air Quality challenge, the datasets, despite both cities being in the same country, are similar, but not exactly the same. City 1 ran two air quality pilots, one with City 2 and one alone.

In the cycling challenge, City 2 focused more on general traffic and road datasets, and City 1 on children, schools and cycling. Somewhat counterintuitively, the same initiative was piloted in both cities and was deemed successful by the SME, city and other stakeholders [PPC2SC07].

7.3 Summary of Metadata Analysis

The analysis of the data sets used in the project show that cities are not consistently or meaningfully differentiating between datasets that are open or closed, and equally they are not consistently differentiating between datasets that they own, and those that are owned by third parties. From this, it appears that cities are using (or attempting to use) the data that they deem likely to be most appropriate, and they are attempting to fit this into an open data context.

It also appears that the comparative importance of promoting the available datasets seems to have reduced in Call 2. This was not a specific decision of the project overall, so must have been driven by decisions of the cities themselves. Both of these observations will be investigated further in the group interview, in order to establish motivation.

The next section moves from the metadata to document analysis.

7.4 Results of Document Analysis

Abductive thematic analysis was performed on the 58 selected project documents, beginning with the top-level categories of access, purpose, permissions, privacy and value. Thirty-one codes were developed in an iterative process, with 27 eventually being used in the results. Extracts may have been edited (or additional text added in parentheses) for clarity where there is insufficient context without it, but in the majority of cases, the text is as it was found. The list of codes can be found in Appendix D.

7.4.1 Access

The fundamental principle that anyone can use open data has several dimensions. Primarily it must be available to use: published and, crucially, discoverable. Across all the groups of documents – open data, pilot process and minutes - themes of internal availability, opening processes and informing the ecosystem emerged. Topics concerning data being of sufficient quality for people to use it, and ideally, conform to standards so it can be interoperable with other data, were also revealed.

7.4.1.1 Availability

The first step for the cities, whether they were already publishing open data or not, was to *"Consider how certain data can be unlocked"* [PPC2SC07] in order that it could be available. However, this is a fairly extensive and unstructured process; "*Think about all the data that might be collected by co-workers and try to make sure the data is collected and stored properly.*" [BLPSC05]. Unsurprisingly, cities found, "*There was no proper process of collecting, storing or sharing the data*" [BLPSC05]. Convincing colleagues was not an easy job, it was a "*challenge to get the data administrators to open up data, [they felt it would be] 'lots of work' and 'scared if wrong*'" [FODW]. Further, as they required specific data to be opened, this was often spread across various departments. A single relationship with one cooperative department prepared to be a test bed was not sufficient. *'[We have a] dependency on multiple parties who can provide a data source"* [PPC4SC07].

The above challenges were compounded where the data was not in fact captured in some analogue or digital form. Once the projects were underway, with a directed aim, cities discovered, *"There's a lot of knowledge inside people's minds… a lot of the things they know are interesting are in people's minds...we try to get it out of their minds...both of the start-ups are working on that"* [City 4]. As noted in the literature review, the majority of data requires documentation (on its original purpose, provenance, even such simple aspects as what fields might mean) and this is no different here, but with an additional complexity around institutional memory. *"You really have to take a dive into documentation to figure that out,"* [City 4]. Ultimately such issues meant

that, even according to themselves, "*City 4 were the 'hiccup' - bottlenecking data opening at the beginning of the process,*" [MPPSC05].

7.4.1.2 Publishing Decisions

When deciding what to open, cities are guided by mandate, *"We also have the argument of 'The law is there, we have to publish it anyway'"* [City 2]. For the project, they initially decided "*to create 6 datasets recommendations per challenge to meet the requirements of the deliverable and prioritise the datasets if possible*" [ODGP] that need to be opened in a shared challenge. However, they soon questioned this approach. *"What about data that city did not think about when mapping data? Maybe companies have better ideas. How do we give them this freedom?"* [SC02] Within the project, once the pilots began, they were guided by the SMEs as to which open data was required rather than working with what is open. "*I'm in contact with the businesses to know what they would like to have for open data*" [City 1].

One of the difficulties the city partners found was having the internal power to dictate what data should be opened. This requires "*local authorities to have the right data management in place and the right mindset*" [SCRT]. This contributed to their struggle with opening the right data. They felt that having a champion in the organisation to work on their behalf was necessary. "*What I'm missing in the doc [WP1_A1.4_Workshop Open Data Guidance Partner Meeting Cambridge] is the identification of an authoritative figure in the organization that can push the publication a bit. To add some pressure*" [SC03]. "*Cities could involve high level representatives to evaluate the possibility to gather the needed data*" [SC05].

7.4.1.3 Discoverability

In terms of access, only one city had an open data platform and associated strategy. This meant that discoverability was currently very low. *'We didn't yet officially launch our open data platform, is was created because we need it in the SCIFI project"* [City 1]. That meant that, "*nobody in the outside world is aware of our open data platform beyond the three companies [we are working with]*" [City 1]. One of the other cities also did not have an open data platform and the other had fewer than 12 datasets [DP2, DP3] At the Smart City Round Table, one (external) participant suggested that, given the discoverability issues, cities were the best consumers for their data. "*Or do they [cities] focus solely on their own data? And how easy is it for citizens and other reusers to find and access the data?"* [SCRT]

7.4.1.4 Standards

The standards referenced among the city are numerous, and they show themselves familiar with standards and initiatives to develop these, but are appropriately cautious. *"Smart Flanders want to run, but we should learn to walk first. Linked open data is too ambitious for SCIFI. It would be ideal, but not a good idea at this point,*" [SCC04]. There is a larger aim to connect the project to a

similar project involving cities in another region: "*Aim is also to connect SCIFI to [similar project] SCORE and increase interoperability, other platforms*." [SC05] On the other hand, there is a wide range of comfort with standards: on the one hand, one city felt that, "*Some externally referenced datasets are in the WMS format, which doesn't seem easy to use in an operational environment"* [BPLSC05] despite the fact that Web Mapping Services is a widely used standard. On the other hand, the final report of one city on their solution stated: "*Solution based on standard to manage the different entities (stadium, parcels, sensors, actuators) Standard issued from different sources: Fiware, schema.org, GSMA Future Plan: Publication and propose this data model as official standard inside the front runner smart city program"* [PPC3SC07]. This is an Open and Agile Smart City (OASC) programme promoting common data models for smart cities. Hence, there is a range of competencies evidenced.

Cities were also driven by the standards adopted by the SMEs. In some cases, this meant that they felt less inclined to publish in a wide variety of formats if one prevailed. "*[We only need to publish in] JSON format, the only format they are interested in"* [City 1]. In other cases, they found the level of data publishing required to be interoperable to be much more arduous than simply opening the data set as is: "*We give access to the data set on the calendar format csv, it's for me open data, but when we talk to the start up, we need to make many modifications to be interoperable"* [City 3].

7.4.1.5 Risks of Access

As well as the risks around personal data, further risks were identified. Essentially, these all fall under the banner headline of political risk, which is obviously of great importance to the city representatives, professionally, and presumably, personally. There was a general risk of misuse. In the Open Data Guidance Package v1, it states, "*To prevent possible misuse of the released data, cities might want to add information when releasing the data. Specifically, if data has a possible political risk."* [ODGP]. Another identified risk was of releasing poor quality data - "*If the quality isn't right we are scared to release it to others*" [FODW] although City 3 also took the stance that, "*The city is not responsible for missing/false data in our datasets*" [SC03].

"*What political responses to the reuse might be necessary, determines for a major part the approach of open data*" [SCRT] is key. Simply putting data in the public domain can send a message to citizens that perceived problems related to the context of the dataset would be addressed by the city. "*[There is a] risk to create expectations [with citizens] on data [outcomes] that can't be delivered,*" [SC05]. Further, data in the public domain can be misinterpreted or interpreted problematically to cause division and social issues. "*Based on the data we published, some groups and some areas were like in a negative image… [the council] are really scared for those conclusions [that people in low income neighbourhoods consume more alcohol]*" [City 4].

One type of data in particular - air quality data - illustrated both these problems, from different sources, one internal to the project and one external. "*Understanding the difficulties in collecting qualitative air quality data and the risk of having people draw wrong conclusions based on open air quality data might have a city think twice before just opening up the data,"* [SCRT]. In City 1, there was insufficient policy context to open the data. "*Too soon to go to the citizens … it was too sensitive let's say...the city doesn't have enough insight into air quality… they are concerned about the reactions from citizens,"* [SME2].

7.4.2 Purpose

The key themes that arose around the concept of purpose were those of guided reuse, internal reuse and the necessity to specify purpose to engage other departments with opening, relationships with reusers and collecting and generating new data.

7.4.2.1 Guided Reuse

Currently, only one city of the four promotes any kind of reuse activity, through "*some social media*" [ODP&R]. Despite this, various views and concerns on the importance of targeted or directed reuse emerged throughout the documents. This particularly came to the forefront when discussing how open data could be sustainable. It was suggested that, "*Cities ...position themselves regarding the use they want to see (e.g. promote and facilitate for the common good)"* [FODW] and that to encourage reuse, the city should, "*Define as a city what you want to get out of it. Example, measure the use of public toilets or shared bicycles*" [FODW].

Cities who actively published open data were the recipients for requests on guidance as to how data should be used, both from citizens and the media. City 4 held a press release on some data sets, and subsequently *"I got called by a journalist and he wanted to know what could be done with the data [we published]"* [City 4]. Yet, *"there's a debate on what extent you should provide thought"* [City 4] that would guide citizens as to potential purposes.

The external voices captured in the documentation (from organisations that are open data intermediaries) echo this. "*Cities need to understand... that the context of the data determines its value. Thus, cities need to embrace 'scenario thinking'. What possible reuse is foreseen ... determines for a major part the approach of open data"* [SCRT] That cities use the approach of consulting their citizens on what issues they find of import that can be addressed with open data is shown in the challenge business case documents, which have a section for displaying the input that has been collected from quadruple helix (business, government, academia and citizenry) sources. The citizen input, particularly, is shown to have impact: the minutes from Steering Committee 03 record that City 3 held a series citizen round tables on their Call 1 devised challenges, as a result of which, *"one challenge moved to 2nd phase and 1 challenge added on watering green areas (energy challenge)"* [SC03].

7.4.2.2 Selected Reusers

Reusers are seen as a key part of this directed opening activity on both sides of the equation - a wider audience assisting cities with identifying possible uses, and a narrower audience being engaged with use after opening. *"How can a city foresee all possible use cases? Not alone and not all...Cities need to cooperate more with businesses and citizens to understand the context of the data"* [SCRT]. "*When we open up a dataset it's always related to a development in the city or a project so… we know exactly who are the stakeholder that might be interested and we try to tell them"* [City 4. This was also seen as financial sustainability strategy at the 'Financing Open Data' Workshop. *"Publish data 'by design' offering the opportunity to reach out to specific stakeholders"* [FODW]. The relationship with users is dominated by the city's opening decisions and outbound activities, but there are also inward approaches, not just from citizens: "*The [national weather service] tried to develop something that looks like the solution [we are] developing, they were interested to get in touch"* [City 4]. For approaches from outside, a more arms-length relationship is preferred, *"It's easy if you can just say 'take a look, they're on our platform"* [City 4].

7.4.2.3 Tracking Reuse

While the majority of occurrences seemed to focus on reusers as the focus of a more human-based relationship beginning at the start of the reuse project, there were also some comments regarding the possibility of a transactionally based relationship that put onus on users. "*Users should negotiate and give feedback on the use of data by clients/citizens"* [FODW] and, "*This [possibility of asking people to register to access data] is also dependant on the platform we're using for publishing. It's useful for analyzing who took the datasets and what's being done with it"* [SC03]. On the other hand, this was seen as having limits in terms of how strictly it was enforced: "*But will you force users to report on their impacts? I'm not in favour of that*" [FODW].

7.4.2.4 Dogfooding

Dogfooding (also, 'eating our own dogfood'), is a colloquial expression to describe occasions when organisations use their own products or services for their internal operations. The ultimate success of the SCIFI project will be whether any of the solutions developed are implemented within either one of the SCIFI cities or another city using its own open data. However, even outwith the larger frame of the project, there is awareness that *"one of the reuses of the data is to the city itself"* [City 2]. "*Re-use of data is an important part of the circular economy. We need to be re-using data not just collecting it"* [SC06].

This varies from very basic uses through various levels of complexity. At the most rudimentary level, *"We hope to stimulate the data awareness and the data policy within the organisation"* [City 1]. One step up from this is internal use to break down data silos and, "*Have people from the municipality use the portal*" [FODW].

However, as cities attempt to utilise their data in more advanced ways there are increasing complications. One is that "*It requires a lot of data to be able to come up with a model*" [PPC4SC07] - perhaps more than a city can easily access, either from their own data or that of third parties. There is also a limit to which of their problems cities can solve, even if they have the data. For example, the de-icing challenge set out to optimise the times and routes for spraying salt on roads when icy conditions prevail. However, the city does not make these decisions, and it does not spray the roads, but contracts this work out - in other words, it is one step removed from possible implementation, even with an ideal solution. "*We might have been looking at the wrong end user; many municipalities only validate the number of dispatches, and leave all de-icing operations to sub-contractors*" [ERPP3]. In this case, the city has subsequently worked with the de-icing subcontractor and multiple other municipalities to attempt to encourage the sub-contractor to utilise the algorithmic model developed.

Defining a specific purpose is also necessary for some cities to be able to engage with data-holding departments in order to extract the data to make it available. "*If you want to open the data you need to involve a specific department, for this department it is more work, if you just opening for opening, for them there is no reward*" [City 3]. Overall, use cases are necessary because it is "*hard to convince the city of the importance of data"* [SC04].

7.4.2.5 Collecting and Generating New Data

Data generation and collection via sensors was seen as desirable, in part because the cities felt this was a painless and cost-effective way to acquire relevant data, especially when it was not obvious which datasets could be used to address a challenge. "*Can collection of data be part of the challenge. Get data to solve that challenge*." [SC02] *"E.g., Solution: data collection in first acceleration round and second acceleration round we use the data to develop solution.*" [SC02] All but one of the solutions utilised sensors. The content is specified, and the location and the structure of data known. "*Data on portal is limited at the moment - we need more. Data will be generated now in accelerator phase- let's publish them as well on the SCIFI portal-> Easier to re-use"* [SC06]. While the plan was that, *"the data from the sensors will be open data"* [City 1] there was also awareness that more work had to go into, *"What the city wants and can do with the inflow of data"* which might require "*smart" filtering of data. -> manual process by the mobility dp*t." [PPC2SC07] and also that the process was more complex and onerous than at first appeared, "*I think the translation into day to day practices and the consequence on budget are not fully considered"* [City 1].

7.4.3 Permissions

The key issues that arose in the category of permissions focused on licensing, data 'ownership', uncertainty around data status, data sharing and charging for data.

7.4.3.1 Licensing

"*Open licensing can be a problem*" [ODGP] and the cities acknowledge that a proliferation of "*unique open data license versions*" [ODGP] is unhelpful. However, in practice, it appears that there are still a variety of licenses, that these do not comply with the standard set within the project and that licenses are not clearly promoted.

The project standard is to, *"Use an open license for open data (Creative Commons zero if possible) to ensure reusability of the data.*" [ODGP]. None of the cities does this. Cities 1 and 3 use CC-BY, City 2 uses the Flemish standard which is a Free License, where the Intellectual Property, is retained, but reuse is allowed, as noted in Chapter 2. City 4 states that data can be used in terms that match CC-0, but do not explicitly use CC-0, stating "*Everyone - private individuals and companies - may reuse the open data of the municipality, as placed on the data platform, in their own applications".* [CDP4] The licensing on the open data is supported by terms in the contract, that state, "*All the data available via the open data to which the Contractor has been granted access to are free to use and reuse for commercial and non- commercial purposes in accordance to the applicable open data licenses*" [CONT]. There is further coverage of non-open data. "*For the purpose of the project leading to the development of open data solutions for Smart Cities, the Community grants the Contractor access to Public Data and/or open data published by the Community and other Cities.*" [CONT]. This terminology is slightly confusing. Public data (public sector information) can contain personal (even sensitive) information, but this should not be published. If it is accessed in a non-published manner, then various GDPR principles and regulations apply. However, what is clear is that there are no further terms on the access, other than a confidentiality agreement.

The cities state that they find it difficult to conform to regulations on mandated opening, so there are data sets that do not have the required permissions attached. "*Many cities have difficulty to apply, to conform to the regulations*" [City 3]. They also acknowledge that, especially across borders, licenses and rights are not standardised. *"It's a fact of life, the SMEs have to cope with the fact that different cities have different rules/regulations on data"* [SC04].

7.4.3.2 Uncertainty Around Openness

The exact licensing of the datasets is not apparent to applicants to the open innovation challenges, as the vast majority of the datasets are not linked to from the application website, but only described (and not necessarily with the name of the dataset as it appears on the site. This is sometimes a function of the translation into English, but other times, a function of unclear naming of datasets). The internal documents do not use licenses to clarify the status of data, using sometimes confusing terminology like *"public (open data)"* [META].

Without data inventories for some of the solutions it is hard to establish exactly what cities or SMEs believe the status of the data is, especially when it is generated or collected, rather than

historically published. However, extracts from the documents show that the Air Quality challenge uses a variety of data, the status of which is not entirely clear to all parties. *"The data through Aircheckr is not open - it was open data [from the Copernicus satellite] but it is improving it… they try to make a sustainable business out of that which is not easy"* [SME2] SME2 runs Aircheckr, so this is in fact not open or even purchased data, but proprietary data. This particular solution was intended to use further data directly from SME3's sensors. The city which ran the project was uncertain whether it was open, "*Measurements [from new sensors] published as open data: Unclear*" [PPC1SC07], however, the SME which wished to use it felt, "*So the data is not really open data*" [SME2], although they did not state what they felt it might be. There is a lack of clarity from the SMEs about the exact status of the data, when asked, 'is that open data?' one answers, *"Uh… I guess so"* [SME1]. They go on,"*I could tell you what is open but for the rest of them maybe open but maybe I have a feeling they are closed"* [SME1]. Ensuring sensor data that was planned to be open was made open was not prioritised: *"We have focused more on what we should do...than opening [sensor] data [for future use]. So we didn't discuss that recently*" [SME1]. When discussing what needed to be achieved during the mid-pilot review, gaining "*insights and help on sensor data as open data?"* was listed as something the cities still needed to achieve [BLPSC05].

There is also some lack of clarity where OpenStreetMap, which as previously noted, entails a sharealike requirement, is utilised. "[*We] Use OpenStreetMap, so completely free*" [SME1]. Discussing which data used in the project is open data, a SME explains, "*the mapping is based on open street map*" [SME2].

7.4.3.3 Data Ownership

Ownership of data that was generated was a further area in which the intent was clear, but the actual implementation was less apparent.

The Contractor will be the sole owner of the intellectual property on the results and outcomes developed by the Contractor during the execution of his project in this Contract (including source code developed or produced in the execution of the Services), and all associated Intellectual Property, with exclusion of the Public Data.

However, 'public data' is defined as that provided by the city: "*The data provided by the Community to the Contractor to execute the Agreement in order to develop the ICT-solution and deliver the Proof Of Concept, as described in the Plan of action."* [CONT] One city had amended their contract to include that, "*the contracting authority reserves the right to publish, under a public reuse license, which specifies the rights and obligations attached to the data, the data resulting from the use of the tool supplied by this contract*" but this was not found in the main version of the contract. Therefore, it is not entirely clear that the sensor data definitely belongs to the city and therefore falls under 'public data'.

In the final review of the pilots all the SMEs were asked, *"How do you assure ownership of the city on the new produced data during the pilot phase?"* [ERPP] Only one SME answered this, however, they provided the link for access rather than any clear governance statement: *"Overview of data & insights for city: <secure URI> credentials: <login/password>"* [ERPP4].

7.4.3.4 Sharing Data

The existence of shared datasets, both at the beginning and the end of the pilots, is also a complicating factor. It is not clear what license is agreed over the shared datasets. As noted above, only open data is covered in the contract. In the other direction (i.e., when sensor data is created and shared back to the city), This is of concern to City 3, which has brought up that this requires clarifying in a new reiteration of contracts [SC07].

7.4.3.5 Elision

Throughout the documents a casual eliding of open data and shared data occurs on several occasions. This is most evident in the records from the Smart City Round Tables, where platforms and use are continually imagined as serving data of all natures, and not particularly privileging open data. *"Do they facilitate a platform where data can be shared or published as open data?"* [SCRT]. "*One example given was that of Jules Le Smart in the Walloon Region in Belgium, where a platform has been developed in a public-private ecosystem that can be used to share (open) data with different organisations that collect data for or about the city."* [SCRT]. This is envisaged as only developing as the volume of data that cities hold increases. *"Despite the challenges cities face nowadays with open data, these might change or increase in the near future. As technology evolves and the amount of data collected grows exponentially, the scenarios of (open) data change with them. Data will shift more and more towards transactions, like now with currencies.*" [SCRT]. One particularly interesting comment was the idea that sticking closely to open data rules was a kind of support framework for data publishing that would become less necessary with experience. *"It depends on the maturity of your organization whether we follow strict rules because it's new, or that the process settles over time and publishing gets less strict."* [SC03]

7.4.3.6 Unacknowledged Sharing

There appears to be a level of unacknowledged sharing, where there is an intention to work with open data, which guides the agreement, but the implementation does not quite meet the standards. "*Most of them [the data sets] are open, some are just internal datasets*." [SME1] This leads to some loose agreements around (re)use where data is effectively shared but under open data terms. "*[The data is] used through API of FIWARE [the project data portal, which contains only open data licenses], I believe the idea is at some point to publish them*" [SME1].

7.4.3.7 Benefits of Sharing

As can be seen in the data tables, two cities ended up sharing data with their SMEs, and both found it very positive, and questioned exactly why open data was actually required. *"More pragmatic and my ideal for me is to give access to [the data] I don't think there is very many [reasons] for opening the data… [I want] for me to be able to give access to the data, to the system to produce new use cases"* [City 3]. "*Is it open data that helps to get to the benefits?*" [City 4]

The key difference perceived was the time involved in creating open data sets, which was problematically long. *"Process to unlock data takes longer so we share first and open later"* [PPC4SC07]. *"One of our start-ups, we sent them some data sets, and we're opening them up right now, and that's interesting - why does it take so long to open up the data and so quick to share, which is basically the same in some way"* [City 4]. These datasets therefore are evidently non-personal, as they can be opened, but the city feels the sharing process is easier, possibly because there is less political and quality pressure when a dataset is not exposed to the wider world. However, they do note the effect that data availability to anyone might have on competitive markets, "*I think you can share data with that party [to develop a solution] but perhaps another party also has a solution like this*". [City 4]

7.4.3.8 Charging for Data

Although charging for data was never considered within the actual SCIFI project, it emerged as something both partners and organisations the partners engaged with had thought about. It arose in the context of value as a way of creating a mechanism for measuring the value of data. "*Everything that you give away for free has no value. So if we could charge for our open data, we could quite easily calculate the benefits of it*" [City 2]. Identifying the 'valuable data' here is of course the issue. In thinking about how open data could be financially sustained, the importance of valuable data grew. "[*We could] make distinction between what for free, what not for free based on quality, frequency of real time data, accuracy"* [FODW]. Another city similarly discriminated between "Realtime vs static, age of data, [specifications] of new datasets, payment for data use" [FODW]The cities were also aware of the technology to enable this. "*[We could] develop smart contracts to monetize valuable data through licensing the right to use it in solutions"* [FODW]. The clause added to City 3's contract regarding the right to exploit and disseminate sensor data, included the term, *"with a view to making public information available free of charge for reuse free of charge or at a cost"* [CONTF].

At the Smart City Open Data Round Table a participant suggested that cities might lead on moving the concept of open data from completely free, to taking into account reasonable costs of creating, maintaining and publishing the data. *"Cities might need to take another role in this ecosystem of data end even consider open data a means that is not always for free."* [SCRT].

7.4.4 Privacy

As open data is not personal data, privacy was rarely formally on the agenda in the SCIFI documents. The 'official' line was present in the documents - *"With different scenarios in mind and technology in place, cities can take the right steps in order to open up data in such a way that it is reusable, qualitative and reliable, easy to find and prevents privacy breaches. Open data by design, thus.*" [SCRT], however many other situations emerged during the actual reuse process that suggested that practice and theory diverged.

There were a few incidences of accidental publishing of data that appeared harmless but, under expert scrutiny in context, was more problematic. *"We found out that we were publishing complaints of citizens on the related topics including an open field that led to data breaches*" [BLPSC05]. *"Privacy issues with some data sets. This was detected by the georef data expert in City 4"* [MPPSC05]. However, there was also at least one occurrence of a dataset used in a solution having been reviewed and found too problematic to publish, and therefore being shared: "*Sports calendar of the city: Remain a closed dataset for security concern according to IT department*" [PPC3SC07].

There is a general awareness that privacy concerns are becoming more of an issue, and that this can not necessarily be fixed by data manipulation processes. *"Some cities that started with 'just releasing open data' now face challenges in the whole process of opening up data, as they have privacy concerns."* [SCRT] *"Anonymizing data might seem enough at first place, but based on anonymized data one can derive patterns that a city might not want users of the data to see."* [SCRT]" (As above, the concern here is not necessarily for individual safety, but for political risk.)

In later documents of the project regarding the pilots themselves, the awareness and practice of issues of consent around data collection (via sensors) are evident. *"The [group of individuals creating the data] was well informed [at all stages of the project], in particular because of the sensitivity around data privacy"* [PPC2SC07]. SC06 minutes state that it is necessary to, "*define clearly purpose/use of data collection. Make sure users give their consent*" and also to *"identify what we can do and where the gaps are concerned to privacy"* [SC06]. However, as previously noted, the documents show a lack of clarity regarding the ownership of the sensor data.

Contractually, regarding personal data, there is a clause that in the standard contract that states, "*For the purpose of the project leading to the development of open data solutions for Smart Cities, the Community grants the Contractor access to Public Data and/or open data published by the Community and other Cities.*" [CONT] It is not entirely clear whether in this case this means 'Public data that is published' (i.e., not personal data) or it means 'Public data' and the 'published' part refers to the open data. While 'public data' can (and frequently does) contain personal data, the separation of the two is not clarified.

7.4.5 Value

Aspects of value that arose in the documents included the value to the city both in its engagement with internal and external audiences; general value to public services: measuring value and the value to the data users (and intermediaries).

7.4.5.1 Value to City - Internal Value

Improving their internal data management processes and data quality is a key purpose of opening data for the cities. "*Opening up as a way of improving the quality (based on the actual process and feedback end users)*" [FODW]. "*We're using the SCIFI programme to get our internal data management in order"* [City 2]. It is an opportunity for city officials and representatives to engage with a resource that is often highly distributed, poorly stored and the potential of which often has low awareness. *"It is a learning experience for us but especially for [the city], because sometimes they don't realise what they want and what data can be used for*" [SME 1].

The city officials also use data as evidence of what they have achieved, and to enable them to present their work more effectively to elected representatives. *"We can go to the city council and say this is how clean the city is based on data"* [City 4].

7.4.5.2 Value to City - City Management

As well as internal day to day efficiency the cities locate value in citizen-facing improved city management. *"What really is important for us as a city is the benefit we're generating...for our citizens"* [City 2]. For instance, by improving waste management services through data they *"want a better guarantee towards the citizens that streets are clean*" [City 4]. Whether the new services are procured or not, *"The impact for us is we have a better view of what's really going on in the city"* [FODW].

7.4.5.3 Open Innovation Success

Despite the entire project focusing on creating pilots to address public sector challenges with open data, there is less focus on the ability of the pilots to actually deliver the value (as opposed to simply completing). "*[Aim is] to get a more open data environment. Curious as to the delivery of the solutions will indeed help solve the challenge"* [SC04]. This is in part because of the complexity of both the projects and the city processes, which involved many more people that worked on the core of the project. "*There was a moment when they [the city] did get stuck because they did not understand what was the final achievement of the project, even if we did filling a milestone document with all the steps"; "Struggled to understand the plan for carrying on…[City4] said 'ok nice but we don't want this because [a previously ignored group of stakeholders] can't use it"* [SME1].

The outcomes of the first set of pilots are not an important measure for this research, but may be of interest for suggesting why there is less emphasis on the value of the solutions themselves. City

3 is running a procurement process for a watering solution, of which their pilot will be one of the tendering companies. City 1 and 2 are reviewing the ongoing costs of the SME cycling solution, to understand whether they can afford to procure such a solution going forward. City 4 developed a solution they could not procure directly for de-icing, but would be interested in procuring if it is developed further in the commercial market. Learnings included "*innovation is fun!*" [PPC4SC07]. City 2 is running a waste challenge in Call 2, and is interested in reviewing SME1's solution (developed by City 4) if they progress to a full procurement. [PPC1SC07] [PPC2SC07] [PPC3SC07] [PPC4SC07].

7.4.5.4 Measuring Value

An entire sub-deliverable of the project, for Open Data Guidance Package v3, focused on how the cities measured the value of open data using key performance indicators from the pilots. In the discussions prior to the creation of the deliverable on value ([City 1] [City 2] [City 3] [City 4] [SME1] [SME2]) it is apparent that most of the cities are either not clear on or not considering the measurement of output or impact to create value. *"Whatever we get out of the pilot, we get out of it, but it's not our main objective"* [City 2]. "*It's hard to make sure there's a correlation between you push open data and something happens"* [City 2]. In some cases, this was because the aim was not clear: "*after 5 months they were not clear about what they want to achieve at the end of the pilot*" [SME1]. City 3 was the exception because there was a clear link between the service and cost reduction: *"For this particular pilot it [measuring value] is easy, for others, it is not so easy"* [City 3]. City 1 made a link between reducing the cost of the project and the ability to continue it, but not the value of the output: *"We are hoping to reduce every cost to a minimum and then it will be a no brainer for the city mayor to continue what we are doing"* [City 1].

This also depends on the type of value that is being sought. There was some lack of conviction that open data enabled all types of value. "*If you want to reach the goal of transparency, open data is probably not the way to go. You can still be transparent in, you know, publishing your reports or your decisions is going to be useful*" [City 2].

7.4.5.5 Value to Data User

There is a clear delineation between the value to the SMEs, who are only concerned with good data management, political risk and happy citizens insofar as that affects any part of the product or service they are trying to offer and therefore potential future revenue streams. *"For the product itself [the value is] how many cities we can get interested in the application"* [SME2]. There were aspects of the product or service that SMEs wished to develop for their own future benefit and marketability *"If you consider the potential of a dashboard that is able to connect citizen complaints to the [solution] that is a super, super kind of feature that would be interesting to a lot of cities"* but the city they are working with cannot integrate the citizen data as it is in a "*bucket*" [SME1]. SME1 can still offer this feature to other cities if they can share data but City 4

cannot accrue this value without *"changing a lot of work and procedures"* [SME1]. *"Normally the client is asking for more and more and more [not fewer features]"* [SME1]

One of the SMEs located value in the creation of employment. *"If your company is a bit successful it creates employment… we have 6 people who do not have a job [in this company] before"* [SME2] This value does not necessarily accrue to either the SME or the city - especially as, in this case, the company is not registered in the same country as the city.

7.4.5.6 Value to Intermediaries

Although this did not emerge to any great extent in the documents, it is worth noting that there are non-city partners in the consortium who value the activity in a different way. One such partner noted that their, *"objective is to spread use cases with open data and open platforms (FIWARE) in the region,"* [SC04].

7.5 Group Interview

The aim of the group interview was to firstly to confirm the accuracy of the findings above and secondly, to seek the cities' input into explaining some of the findings. The complete topic guide for the interview can be found in Appendix C.

7.5.1 Amount and Availability of Data

To begin this discussion, the cities were shown exhibits. These were the tables in the metadata analysis. The response focused around the table 'Total Number of Datasets Available by Category and Call' (Table 35). In terms of the comparative number of datasets per call, the initial response was sceptical, calling it *"counter-intuitive".* City 3 felt the disparity of datasets could be due to the different stage in the cycle of the second civic accelerator. The process for determining the tables was presented. Subsequently, a variety of motivations for gathering less data were surfaced. One reason was that the city was aware that they simply did not have appropriate data for their challenge, so they focused more on selecting the most important challenges for their citizens. They noted that in the first call, "*we didn't know what to expect so we took a broad view of what may have been interesting. We noticed…by talking to SMEs the solution they were proposing was more important…So we were waiting to be guided."* Another reason was simple pressures of time and resource: *"Maybe we didn't do the exercise of discovering our data as well as we should."*

On the other hand. the number of categories of ownership was less surprising to the cities. One city in particular immediately agreed with the analysis, noting that they had. "*discovered quickly we need a mix of different data and parties to create value from [our] data*" For the second call challenges this effect was even more marked, and, "*the interesting data sets for our challenges*

are not ours…for example, we know there is a bike sharing solution…the data's really interesting but we know they're not going to release the data".

7.5.2 Sensor Data

The findings on sensor data that were shared were the opacity around consent to use, the lack of clarity over access and the potential for being personal data. In all the cases where sensor data was involved, there was a gap between what was agreed in the documents regarding use of and access to sensor data, and what had actually happened. The cities still leant towards relying on what was stated as happening, rather than what was actually happening. For instance, City 3 claimed, *"in my case all data is published on the [FIWARE] data portal*." However, this is not visibly so. Similarly, City 4 stated that they have contractually agreed with one of their pilot SMEs that the sensor data would be published on FIWARE. However, again, a check showed this had not happened at the time of writing, 6 months after the end of the pilot. City 1 had received the log in to the sensor data generated in one of their projects (as shown in ERPP4), which was still physically with the SME. They were uncertain about the terms of the contract, "*I'll have to double check the contract but I think we included that we must have access*".

Both Cities 1 and 3 said they were not really able to talk about how useful the sensor data was. In one case this was because, *"the data is still with [the SME] and we are not really close with its role"*. In the other case, the city admitted they were, *'not sure to what extent we can use it".* This is an interesting insight to what might happen when an organisation – especially a public sector one in a smart city capacity – owns data which is not in its own digital domain.

7.5.3 Personal Data

The initial topics for discussion on personal data presented to the cities were how personal data might sometime exist within less sensitive datasets, and the potential for any data to become personal via use. City 4's citizen report, which they had been unable to share with SME1 as it contained so much personally identifying data, had informed their ongoing approach to citizen report data collecting. They eventually resolved the issue in the pilot by extracting and retaining the personal data without affecting the content, and this led to them, "*taking a wider perspective on all citizen reports in future [in order to ensure we are not collecting personal data unnecessarily]".* The system had previously included an open text field into which citizens wrote personally identifying information and this was now removed.

City 3 experienced an issue where the IT department did not want to publish a dataset – the calendar of events of the playing fields – that did not seem personal. This was because, *"they didn't want to publish calendar in case of bad feeling [towards a team that was playing] and make an attack."* In other words, this transpired to be a public order and safety issue.

The cycling pilot generated sensor data that was "*all about the personal data*" and "*they [the SME] have collected all the consents"*. The resulting data is never intended to be opened. However, the consents were designed and are stored by the SME. The city is not clear on what the consents cover and it may well be that the consents are limited to this particular purpose (especially as children were involved in the pilot). It should be noted that consent is not the only lawful basis for processing personal data, and as a local authority the cities could well use the bases of legitimate interest or carrying out a public task, however, consent was the only lawful basis that arose.

7.5.4 Data Sharing

The key findings on data sharing that were presented back to the cities were that it appeared that they had used it not only when opening data was impossible (because of privacy issues) but also where opening data had been arduous. There was also inconsistency over how this is managed in contracts.

The data sharing was driven in part by the process of the pilots. The time pressure was not only that the data had to be available to the SMEs within a six-month timescale, but also that *"if you're doing it [opening data] for six months only that's a big investment if you're doing it for the pilot. It's easier, it's cheaper [to share the data]*". They were very aware that, *"if it's open data, you need to make sure there's governance behind it, that there's a model behind it*." However, they did not see data sharing as an end game, but as part of the process towards opening data, "*the trajectory for your data, it has to move with the pilots, so it has a lifecycle of open data like we have of a product*."

Given the greater use of data sharing the cities were looking to change the contracts going forward in order to accommodate this. Key aspects that they were considering that needed to be contractually managed were purpose, "*if we share data we need to have some kind of contract on what they [the SMEs] can do with the data*"; and time limitation, "*maybe at the end of the pilot they [the SMEs] destroy it."* The other role for the data contract was to ensure that city did not give up control of or responsibility for the data. "*You still have to make sure you are responsible for the data, that it's not going to appear on the interne*t."

An area on which the cities found themselves in debate was the role played by the maturity of a city's open data processes and policies in whether data was shared or opened. City 4 suggested that if a city had just begun their open data journey they "*don't know which buttons to press [to open data*" so sharing might have a more important interim role. City 1 argued that, regardless of the stage of maturity, opening data would always be a financial consideration. "*It's a question of return on investment. Even if you're a very mature city it might be very costly to open up data*

sets." Subsequent discussion suggested it might be the dataset itself that influenced the ease and cost – some might be comparatively easy to open, others comparatively expensive.

7.6 Summary

In this chapter I have presented the results of analysis of three sets of data from the SCIFI case study:

This evidence suggests that the open data movement has had great success in creating a culture of open data, where expectations of the use of data and the understanding of theoretical benefits is high. However, this is accompanied by a range of contradictions in practice – not all of which are simply associated with the maturity and experience in data of the cities. Undertaking the project was possible because the cities believed that open data could solve public sector problems (even if this was only their own internal data management). However, the actual use of open data in a way that conforms to the open definition or other definitions of open data has not happened – because this is still largely aspirational rather than practical.

The version of open data presented in the findings above could well be described using the term "*fuzzy open data*" (Huber, Wainright and Rentoccini,2018). The cities were evidently also engaging in ad hoc data sharing, which was not accounted for in the contracts. They saw this as part of the process of opening – an interim stage which would ultimately speed up the process. The generation of new data through sensors was seen as an effective way to produce data which was fit for purpose, rather than historical data that might or might not be useful. Subsequently, issues around the ownership and access to sensor data – as well as associated issues of privacy – arose, but were not clarified.

Below, the key variances in the results from the framework defined in Chapter 4 are shown.

Figure 13 Variations from the Access Framework Definition

Figure 14 Variations from the Purpose Framework Definition

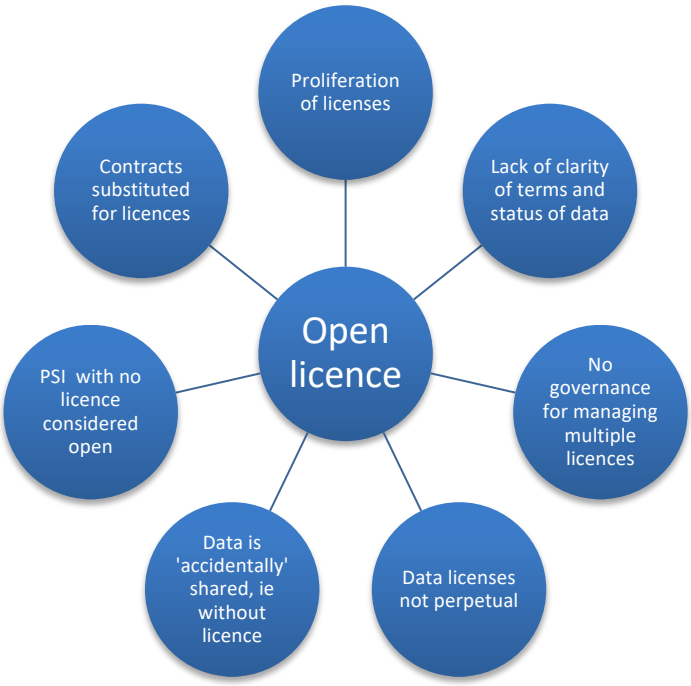

Figure 15 Variations from the Permission Framework Definition

Figure 16 Variation from Privacy Framework Definition

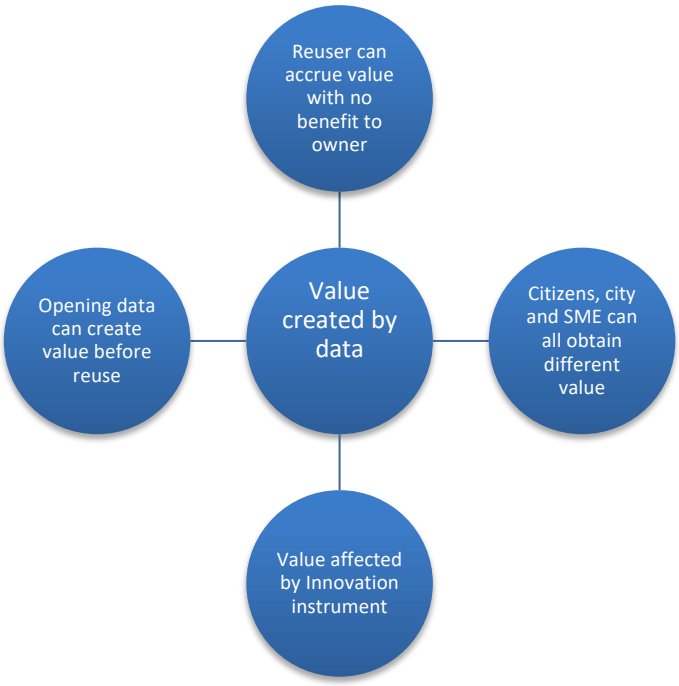

Figure 17 Variation from Value Framework Definition

The next chapter discusses interpretations and implications of these results and synthesises lessons learnt.

Chapter 8 Discussion - How does the use of open data in open innovation in practice vary from the previously defined framework?

The issue this research focuses on is whether the way in which open data is used aligns with the paradigmatic expectations set out by legal and social frameworks, or whether it is at variance with these. An understanding of this is important for both individuals and organisations working in the area, as well as for policy makers.

In the previous section I presented my results for the research question, 'How is open data being used in practice?' via a case study of a European public sector open innovation with open data project.

The results indicate that open data in practice does not match the framework of open data use in theory over a number of dimensions. Although cities are aware of how open data is defined and how to use it, they find enacting this in reality a challenge, both for many of the reasons previously noted in the literature and also some less commonly cited ones, such as political risk. These complexities of opening also are problematic for a timeline: they cannot wait until the data is ready, vetted and of an appropriate quality, so they adopt, sometimes knowingly and sometimes less consciously, a workable version of open data that enables action. Sometimes this presents as an edge case of a compliant open data strategy, for instance user-led opening. (Key Finding 1).Sometimes the Open Data Charter goal of 'open by default' presents more of a problem for the cities - for instance access by anyone and the reputational or political risk this incurs, even where there is no legal or technical risk. (Key Finding 2)

In some instances, what the cities do conflicts with the fully compliant definition of open data, and while the cities are finding ways to make it work, it is exposing them by not having agreed, understood frameworks in place, such as when they share data directly *en route* to opening it. Sometimes their activities conflict with the paradigmatic version because legal and technical rules and concerns have eroded the paradigm. (Key Finding 3.)Lastly, cities are opening open data for a value that is specific to themselves, and is not the same as the value accrued by users, or third parties, (such as tax), and this dictates their priorities.

It is possible to see these results as another way of understanding the barriers to successful open data publication and use, and to continue to attempt to create normative research on what strategies suppliers and users should adopt in order to create value with open data while also aligning themselves with 'model' open data theory. However, a plausible assertion is that it is simply not possible in practice to create meaningful benefit that justifies the cost and time of provision from open data as it operates in theory. This is not only because of the constraints on the activities of the cities, but because of constraints on the data. Legal, economic, political and

technical pressures have combined to erode the possibilities of open data. The users in this study are adapting to these, and evolving a version of open data that reflects what is possible in practice.

8.1 Users and Purpose

The main economic question - one that was hyped by such purported aphorisms as 'data is the new oil' and the McKinsey $3trillion (Manyika et al, 2013) - that hangs over open data is 'why have we not seen these economic impacts?' The literature takes two routes in answering this. The first is that we do not have the capability to measure open data use or impact in a way that proves causality (Lammerhirt and Brandesescu, 2019). The second is that open data requires, "a shift from supply-to user-driven open data provision" (van Loenen, 2017).

Opening data without knowing the purpose is certainly akin to throwing a handful of darts at a dartboard and hoping one hits the bullseye. There are ways to reduce the number of darts that miss the bullseye. One is to open a substantial number of datasets around a single theme, such as geodata or transport, ensuring greater use via ecosystem or network effects. Many of the more successful open data portals are single theme data portals (Koesten, Walker and Simperl, 2020). Identifying a purpose with an external organisation also reduces the number of wandering darts, which, as the cities noted in the interview, are essentially time consuming and expensive. However, having a purpose means having to find the right data. User-driven open data provision does not conflict it anyway with 'for any purpose', but it does open the door to the scenarios seen in the research, where the specific needs of a target user may actually be data that cannot entirely be opened, as it is too sensitive. What, then is the option? One choice is to end the relationship once it has been decided that a sufficiently significant amount of the required data cannot be made open. But this must then be done after some time, effort and resource has been invested in the relationship. (In fact, this happened to City 4 in the second Call – after they had selected a SME to work with, but before the contract was signed). Similarly, if data is unlikely to be used in another capacity apart from the purpose it was opened for, the cities question whether this justifies opening it all. Relationships become crucial - potential reusers engaged with the supplier may be able to prioritise their data needs over the needs of others. (Key Finding 4). By the second call, three out of four of the cities felt that discovery of data before the call is not as important as identifying the important problems to solve and finding companies with the right technologies and capabilities to work on those problems.

This idea that the relationship was more important than the open data is also supported by the example of a solution co-piloted by two cities. The Open Data Guidance Package v1 recommends, *"If cities work together in one challenge, it is recommended they analyse the datasets with similar data and strive to harmonize the datasets as much as possible for solution development."* In the

cycling challenge, City 2 focused more on general traffic and road datasets, and City 1 on children, schools and cycling, as can be seen in the tables. Somewhat counterintuitively, the same initiative was piloted in both cities and was successful. This suggests it was the engagement with the cities rather than existing open data that was the success factor. In the Waste challenge in Call 1, Cities 3 and 4, despite being in different countries, manage to mirror datasets quite well. However, only City 4 selected a SME to work with, as City 3 did not find an approach they liked. Dataset harmonisation, therefore, is no guarantee of solution harmonization.

On the whole, the cities find opening data without a purpose difficult for a number of primarily internal and political reasons. This is demonstrated by the zero or limited number of published data sets available prior to the project from the cities that do not have a strategy for engaging with users. There is more incentive, not only for the city information managers but also for all other city workers, to open data if there is a purpose, a reward, a tangible direction in which to take it. Also, opening data sets without a purpose or target user invites the question, 'what should this be used for?' which essentially requires the cities to manufacture a use reason. Hence, it is more efficient to have a genuine use rationale. (Key Finding 5)

How otherwise should they decide what to open? Nationally, governments open data sets because they are mandated to (Carrara, Radu and Vollers, 2017). Mandating works less well at the municipal level - as demonstrated by cities 1 and 3. Even where openness is mandated, for instance in France, there are issues of monitoring and enforcement. Another option is to decide what comprises 'high value data sets'. An approach to this is to identify what other cities have opened. Even then it is difficult to follow the logic to open unless something has been achieved with the opened data (and that achievement is of value to the holder of the data). The EU is defining a "List of High Value Datasets to be made Available by the Member States under the PSI-Directive" but is this relevant to municipal as well as national authorities? Ultimately, there are increasingly few alternatives to user-driven opening strategies. However, these can be expensive and time consuming, so they need to create results.

Once purposes are in play, the open data trajectory is more dynamic. Explicitly stating a purpose defines a potential use. This has two effects. The first is, the initial conditions for sharing, rather than opening data, are fulfilled. Secondly, once the desired use is known, it is easier for the suppliers to understand what data to focus on (to the exclusion of other potential data sets). Thus is created the potential for a gentle incline away from open data, that may start with "*we know exactly who are the stakeholder that might be interested*" and is part of a continuum that ends with, "*selling data with smart contracts*". Considerations of this kind of outright 'pricing' of open data was restrained to the 'potential' rather than 'actual' list of activities.

8.2 Data Availability

The four datasets belonging to the city are noted as being 'available soon'. City 2, challenge C9. However, there is no guarantee that data that is currently not open, will actually be open at any point, either for political, legal, technical or administrative reasons.

This requires robust internal mechanisms for engaging the right parties in preparing and approving datasets for release (Kankanhalli, Zuiderwijk and Kumar Tayi, 2017)

Despite this being an open innovation with open data project, City 1 promoted no open data at all in Call 2. Their experience in Call 1 had told them that their most useful role was identifying the important challenges for the city, and that developing the datasets was a more collaborative act. In Call 2, City 2 also hugely reduced the effort put into presenting datasets. Although some of the Call 1 datasets might have been appropriate, it is not possible to assume that potential applicants would find them in the previous call information. Fundamentally, City 1 and City 2 focused on the challenges because they did not have the data, but also the data should not be a constraint on the innovation.

In general, the cities learnt from the first civic accelerator that preparation of datasets is not a good investment if the end solutions - which can vary widely - do not use them. Conversely, moving the cost of data preparation (or collection and generation) onto the SME is a more financially attractive strategy.

It is an aphorism that data becomes more valuable when it is combined, and there is a real benefit in being able to use third party data as well as that of the city - all of the three solutions that have data inventories use this. What is more challenging is when the third party datasets are not open, and they are still promoted. City 1 set itself a challenge by specifying a dataset owned by the national government, and which was closed. In this case, it was not needed as the pilot was never run. This might be an example of just 'thinking big' - as, for instance, when they listed Waze traffic sets in their metadata list, but it may also be an example of being too vague about what datasets it is *possible* to use, rather than *productive* to use.

Promoting data from other owners without direct consultation is a risky strategy. If it is necessary to rely on the data sets of others, then it is necessary to rely on them also being open, unless there is a mechanism by which a request to share can be made. Even if it is open data it runs the risk of being withdrawn (without warning, as there is no requirement for a way to contact users). Where it is not open (as in BC7 and the accident data) this is even more risky, as here, where it was neither opened nor shared. Once engagement with other owners is necessary, then some of the friction reducing benefits of open data are removed. (Key Finding 6)

Overall it can be seen that only around half the data is actually owned by the cities and open for the beginning of the digital innovation contests and even the pilots. Even if the cities feel that

either all the data is available, or will be available, or they simply don't want to discuss the complexities of this on a challenge page, this does not accurately reflect the situation for the inbound participants (the SMEs).

The cities were not agreed on whether maturity affected the ability to open data effectively. On the one hand a more mature city might have better processes, policies and political support, on the other hand, the cost of opening data may not justify itself unless there is a clear path to a return on investment via use, whatever the maturity of the city's open data policy. These are substantial financial and political impedances to the 'open by default' concept.

8.3 Data Sharing

Matching licenses reduce friction (Dodds, 2016). However, as noted above, there are multiple licenses involved, proprietary data (including proprietary data that is an enhancement of originally open data), sensor data that is planned to be open, a certain amount of data that is open seemingly by agreement rather than license, and a variety of open (including open Street Map).

City 3 utilised a dataset that was 'public but restricted in some attributes' (and ended up being shared). This is a difficult situation for data publishers. On the one hand, this seems to be an example of a situation in which problematic fields can be removed from a data set before publication. On the other hand, it may be that the removal of those fields degrades the data to the point at which it is useless for the intended purpose (without consulting the IT department this cannot be confirmed either way). Keeping the data in a kind of limbo of 'open but restricted' may introduce greater friction (and possible accidental release of the restricted fields) than simply committing the data to a clear sharing license.

Essentially, therefore, opening may not reduce friction by as much as anticipated. Substantial friction has also been reintroduced to the process by the open innovation mechanisms. This provoked one city to interrogate whether it is actually the fact the data is open that offers the benefits. They noted, insightfully, that sharing data with only one company has the potential to cause a failure in the competitive market. While sharing data with one company in order to shortcut opening issues certainly may contribute to this, a structured data sharing programme where many users who meet the preconditions can still continue to access the data, should not. Data sharing on the road to full opening (where that is actually possible) can also help provide a model and a use case for further uptake once it is open. As emerged in the group interview, the cities felt that 'data dumping' on SMEs was faster and less expensive, and was a route to preparing data for possible opening, part of a "*lifecycle*" for open data. However, they interrogated why sharing should be so much faster than open data. The limitations of the contract, and the fact words 'data processors' and 'data controllers' did not occur in the

documents analysed, suggests that reduced, and possibly insufficient governance is the underlying reason. (Key Finding 7).

The embracing of sharing data rather than opening it leads to an erosion of open data's protected status, an acknowledgement that other methods could possibly achieve what open data can achieve, perhaps more efficiently, and perhaps with a similar (or even lesser) amount of work, because of the reduced risk of data in the public realm. *"I could tell you what is open but for the rest of them maybe open but may I have a feeling they are closed"* [SME1]. There was a substantial lack of concern, especially from the SMEs, as to exactly how they were accessing the data - this may be because they felt protected by the contract. This also may have reduced the reluctance to use because of varying contracts as identified by van Loenen, Janssen and Welle Donker (2012).

Regulating the use of data that is 'a bit open' is difficult. There are clearly a substantial number of tools in place that are intended to do this, for instance in SCIFI, the contracts, and in general, laws around data protection. However, attempting to identify and comply with a range of separate laws in order to do one thing - share the data - is onerous. This would suggest there is room for guidelines on how 'open' data might be shared, for instance, before opening, or how sections of the data which looked like it could be opened, but after modification could not, could work alongside more conventionally open data. (Key Finding 8)

Data sharing is a way to share the cost of cleaning and preparing data for (possible) opening. In 'co-creating' data sets the SME takes on some of the expense and also contributes its data skills; it can also participate in decisions about quality tradeoffs. Data sets used in context are likely to reveal issues that simply looking at them will not reveal, as discovered by participants at the hackathons attended by Thornham and Gomez-Cruz (2016).

Data sharing can also be a cautious way to manage - either short term or long term - potentially risky data. There is particular risk created by the ID location and time stamp that may be used by the SME working on City 3's parking challenge in Call 2. This may, if combined with information such as parking spaces for disabled people and other located based data, start to create personal data. Opening this data will therefore depend on how it is used in the final product. Yet once it is opened, it could be used in any way and can no longer be controlled by the city.

8.4 Sensors and Privacy

As seen in the literature review, the GDPR has already been the cause of debate around the finer points of open data, such as the existence of data that can be published, but not processed (Hanecek, 2017). Sensor data and privacy convolutes the issue of open data further. On the one hand, the results suggest that sensor data is popular, even preferred by cities (especially City 1). It

does away with the complexities of trying to establish how pre-existing datasets can be 'remixed' to create new products. It bypasses issues of collation and preparation in favour of data that is automatically created and natively digital. On the other hand, although the contracts (both the original and the French versions) stated that the cities needed access to the sensor data in practice, none of the cities actually had control of the data (or equivalent access) or detailed information on the parameters of use going forward. This is part of a larger smart city issue regarding the technical and legal administration of situations where third parties control public authority data.

Where the sensor data is not, for instance, CCTV, or other clearly personal data - where it may seem innocuous - cities are committing to open the data. This may not be practical, given the current separation of the cities and their sensor data. It also emerges a further set of considerations that need to be addressed. Firstly, at a higher level, the set of possible reuses of open data can only be restricted in terms of what is legal and what is not. As van Zoonen (2016) points out, the combination of non-personal sensor data (such as the level of fullness of a wastebin) with service purposes (when to empty the bin) suggests this is a safe area for policy and government, as "*data-abuse*" is unlikely to affect individual citizens. However, the data may of course easily be personal already, potentially because all or most sensor data should be treated as personal per the Mauritius Agreement (2014) or highly sensitive by default, per Kroger (2019).

This is an example of how privacy is creating conflict at the heart of open data, and may even fundamentally make the vast majority of open data, as it stands today, untenable. (Key Finding 9)

8.5 Value

The macroeconomic assumption of value of open data has generally told a story where (a). any chargeable cost is limited to the marginal cost of reproduction (virtually nil, thanks to digital technologies such as the web) and (b). value is obtained by a somewhat circuitous route of more businesses gaining more revenue, therefore creating more employment and more tax. (This story obviously changes somewhat for private businesses opening data, where increased efficiencies reduce costs, or improved products and services increase profit, but this is a tiny percentage of open data). This macro narrative of success as employment was still present, as related by one of the SMEs, but this is not relevant to the city.

In this case study, the value is a much more direct, two-sided calculation. While the businesses still play their part in the equation, with the aim of increasing revenues from new markets or products, the cities are potentially benefiting from the new product or service that can serve their citizens. This is because of the use of open innovation. Unsurprisingly, the cities have a greater engagement in the process of creating (and failing to create) value than in situations where only

outbound open innovation (publishing) is involved. However, they can obtain for themselves two kinds of value, both of which are potentially much less valuable in simple outbound activities.

The value for the city is "*the benefit we're generating...for our citizens"* but this may not only be derived from the output or impact of the pilot. "*Whatever we get out of the pilot, we get out of it, but it's not our main objective.*" Here their objective is improving data practices. They can derive benefit from the process of opening, whether the actual open data has any value or not, either to the city or the SME. In fact, the cities all struggled with articulating the externally facing benefits of the process of open innovation with open data. This is not because there are none. It is partly because they are not set up to measure or track the benefits. However, the learning would work equally well for the city whether they are successful or not in innovating. (If they focus the opening of specific datasets for a specific purpose they will also learn more about their own data gaps than if they just open what is available.) This value does not transmute into value to the organisation or individual that used it.

As an example, consider the situation of the SME which wished to develop the dashboard that integrated citizen complaints. Although the city was not able to supply the data within the time frame, they were able to gain the learning on the importance of that data and what needed to be done to gather it. In this scenario, they obtained the learning, and later on, they may still be able to obtain the service, if it gets developed with another city's data. The SME only gains value if they can actually develop and then sell the integration.

On the other hand, for a number of operational reasons, some successfully developed pilots were unable to be procured by the city. In the case of the optimised de-icing scheduling and routing pilot, the city was dependent on the solution being adopted by the de-icing contractor, and this was only fiscally prudent should the other customers of the contractor also wish to improve their de-icing prioritisation. In the same city, the workers on the ground became the bottleneck to using the new system. However, value can still accrue to the SME who, armed with the knowledge of possible implementation barriers, can aim to sell their service elsewhere.

Coupled open innovation is creating (possibly) more work for cities but also almost certainly more value, due to the increased number of value opportunities, including learning and data preparation. It also creates more value for the SMEs, both through the access to data that is not (yet) open and through the closer relationship to the data holder. Therefore, the instrument of open innovation should also be considered as a locus of value. The benefit, which could hypothetically also be that of democracy, transparency, accountability or evidence-based policy, is in fact divided into the benefit to the supplier and the benefit to the user, both of which are dependent on the mechanism that brings them together. (Key Finding 10)

8.6 Limitations

While the research looked at a particular case study in depth, the generalisability of the results is limited by the focus on one particular open innovation project and four northern European cities. However, the findings of the various issues that the cities are grappling with are in line and build on a previous body of work on barriers to use of open data (Conradie and Choenni,2014; Martin et al, 2014; Barry and Bannister, 2014; Zuiderwijk et al, 2012). It is therefore highly plausible that other municipalities engaging in similar civic innovation projects will be engaging with open data in a comparable fashion.

8.7 Summary

It is a relatively clear cut job to define the parameters of what constitutes open data on paper. Once this is transposed to the context of directed reuse– creating pilots to address problems – there is less space for privileging open data, and constraints and pressures come into play. The time constraints of the open innovation pilots also create the conditions for data sharing as part of a trajectory to open data after the pilots, rather than investing in an extensive (and expensive) preparation period before. New sensor technologies and the collection of city-owned data by third parties adds another layer of complexity.

Key findings of this research are:

1. The cities use a variety of approaches to opening data, not just a bottom up or top down strategy, depending on what is needed. This has implications for scalability and re-use. The cities are not opening data until they know what they should open (and then sometimes they cannot open);
2. The cities have key political risks of opening data that limit the opening of data that is otherwise technically and legally compliant;
3. The cities process for making data available sometimes deviates from open data best practice, and frequently presents as edge cases;
4. Relationships naturally grow up around user-driven data provision. The close relationship perhaps leads to increased 'fuzziness' about rights and licenses, what is covered by the contract and what needs separate specification;
5. Having a specific purpose for opening data justifies the work involved in making it available, and successful use offers a way to demonstrate the value of the work involved;
6. The cities are not differentiating between open data, publicly available data, private data until the practicalities of the situation force them to. In other words, their vision for data they would wish to use includes both data they control, and data they don't (data that is owned by third parties). This is a constraint on their activities;

7. They are finding sharing the data - providing direct access to data to named parties for a specific purpose - much easier than opening it in the short (and perhaps medium) term. However, this appears to have insufficient governance around it.
8. There may always be 'edge cases' of open data - user-led opening, data which might need to be understood in use before it is fully opened, or only part of which can be fully opened. The previewing of this data by, for instance, initial sharing in practice or making only the metadata open, might be a viable approach here.
9. The interface of open data, automatically generated data, the physical storage of sensor data, the GDPR and technical capacity for re-identification has severely limited the range of data that can be opened;
10. Value is explicitly two sided (this may be of use in trying to measure the impact of open data) and affected by the open innovation instrument.

The research contributes to a clearer understanding of the use of open data in practice. In some ways, data publishers are still struggling with issues that have existed for nearly a decade. But in this open innovation context they are solving some of those issues by sharing data. This is facilitated by the existing relationship with the open data user. However, they are effectively doing this by sharing data as 'non-compliant open data' rather than utilising any of the frameworks from data sharing. If cities are operating in this way, two things are vital: providing assistance to ensure this is efficient, and providing guidance to ensure this is legally compliant. These results should be taken into account when considering the resilience of the open data framework. Given these findings, this study sets out to reconceptualise the open data framework to reflect public sector open innovation praxis and address the question, 'How should the open data framework be revised to reflect practice?' (Research Question 3). This question is vital, because it is only by understanding how open data operates in actuality that appropriate legal, technical, social and political structures can be put in place.

Chapter 9 Results - How can comparison of other types of public and private data sharing arrangements inform the framework defined in RQ 1 so it more accurately reflects open data for open innovation as found in practice?

The data sharing literature is extremely broad. It encompasses a wide variety of activities sector-specific, government-regulated data sharing such as transport in the Netherlands as seen in Klievink, van der Voort and Veeneman (2018) and more commercial sharing such as described by Schwabe (2019). This chapter takes an overview of commonalities, attitudes and solutions that have been posited to the challenges of data sharing.

9.1 Data Sharing Versus Open Data

Identifying those aspects that could illuminate effective approaches to improving the effectiveness of open data for open innovation is challenging. This is both helped and hindered by the inclusion of open data as a form of data sharing by some authors (Schomm, Stahl and Vossen, 2013; Eschenfelder and Johnson, 2014; Klievink, van der Voort and Veeneman, 2018). Hardinges and Wells (2019) express open data as a version of data sharing rather than its own, discrete activity: "*There are examples of good practice—the publication of government open data here; a cross-industry data sharing platform there*." Conversely, Carnelley et al (2017) situate open data as separate from, but not in opposition to, data sharing, *"strong positive complementarity between the open data market and data marketplaces"*, and a potential virtuous cycle which may reinforce their mutual development. Eschenfelder and Johnson (2014)'s research found similarities to the situation of the cities, writing, "*Our results illuminate complexities hidden within the term open data*." In their case, some of the complexities resided in terminology and the difference between 'restriction' and 'control' - for instance, users may have been asked to register to access data, but this was for statistical, not access restriction, purposes. They also noted that many repositories that considered themselves open did not permit commercial reuse.

9.2 Data Sharing and the Framework of Open Data for Open Innovation

Out of this wide spectrum of data sharing activities, of interest are those instances that either enable sharing between specific parties (as in the open innovation undertaken by the cities) or those which create what Richter and Slowinski (2019) designate the *"data sharing club,"* which any party which fulfils a specific requirement can join. Private data sharing arrangements are not included.

9.3 Types of Data Sharing

Data trusts are legal structures that *"provide independent stewardship of data for the benefit of a group of organizations or people"* (Hardinges and Wells, 2019). (Stalla-Bourdillon, Wintour and Carmichael (2019) work focuses on data foundations, a Channel Island based legal alternative to a trust.) The majority of published research on data trusts is around either the use of the legal format of a trust (e.g. Hardinges and Wells, 2019, Stalla-Bourdillon, Wintour and Carmichael, 2019) or using the concept of trust to drive data sharing (e.g. O'Hara, 2019; Mulgan and Straub, 2019).

Grossman et al (2017) define a data commons as *"cyberinfrastructure that collocates data, storage, and computing infrastructure with commonly used tools for analyzing and sharing data to create an interoperable resource for the research community"*. A parallel is drawn between data sharing communities and natural resource sharing communities (Fisher and Fortmann, 2010). Their model is almost entirely used in research formats rather than commercial or public sector open innovation, although Eschenfelder and Johnson (2014) allow for this use.

Susha, Janssen and Verhulst (2017a) define data collaboratives as *"cross-sector (and public-private) collaboration initiatives aimed at data collection, sharing, or processing for the purpose of addressing a societal challenge"*. In aiming to help solve complex societal problems they are a "*tool of modern public governance*" (Klievink, van der Voort and Veeneman, 2018). In this construction, they also aim to address one of the main challenges the open data movement has faced to date – achieving high-impact results and solving pressing societal problems with data. Klievink, van der Voort and Veeneman (2018) consequently place them at the interface of open governance, data-driven innovation and public private partnerships. This last point is of particular import to data collaboratives: private companies are now generating and collecting vast amounts of data (especially those involved with transport, utilities or similar) and are likely to have superior analytical capabilities to government (Susha, Janssen and Verhulst (2017).

The selling of data is not new: Dun & Bradstreet has been creating and selling data on companies since 1841. Carnelley et al (2017) suggest that while data marketplaces can still be simple 'online stores' for data, they are online evolving into a more sophisticated intermediary and infrastructural role. However, the two- (or more) sided aspect of a marketplace is required for other authors, where it is insufficient to simply have vendors (owners of data who are willing to sell) and it is necessary to have trade (i.e. engagement from both sellers and purchasers). The model can be one-to-one, many-to-one, one-to-many or many-to-many (Koutroumpis, Leiponen and Thomas, 2017; Schomm, Stahl and Vossen, 2013).

9.4 Data Sharing Concept Matrix

The table below follows Webster and Watson (2002) in identifying key concepts from the literature (here, identified deductively using the categories in the framework developed in Chapter 4).

Table 44 Concept Matrix for Literature Review

Article	Access	Purpose	Permission	Privacy	Value
Hardinges and Wells (2019)	Beneficiaries; "Fiduciary duty" about best use; questions about regulation of deposit; unexpected harms; data pooled	To share data well, need to start from a purpose	Social license (trust); legal framework of data stewardship	Privacy Enhancing Technologies (PETs)	Trustors, trustees and beneficiaries; encouraging involvement by the private sector, direct and indirect
Young et al (2019)	Initial access to customised synthetic data; legal and technical harmonisation to allow cross-dataset linking; central warehouse	Various	Memberships (for ongoing compliance and research purposes) modular data sharing for ad hoc use; third party public private trust; structured data use agreements	Synthetic data, strong data governance;	Fairer algorithms without availability bias
Stalla-Bourdillon, Wintour and Carmichael (2019)	No centralised hosting of raw data; role-based and purpose-based access control; data added ad hoc	Clearly stated objectives, workflows and safeguards; regulated decision-making processes	"Rulebook" for whole lifecycle of data; Independent governance body; flexible membership	Hard and soft PETs	innovative data-driven processes to generate socio-cultural and economic benefits for citizens, the public sector and business
Hall and Pesenti (2017)	Standardised, repeatable	Purpose and analytics negotiated	Trustee brokers access, purpose and storage agreements and conditions of commercial value (contract)	Trusted advisor for GDPR	Mutually beneficial
O'Hara (2019)	Could be mediated or limited. Data is held in original data stores by original owners	"Terms and conditions" for sharing; audit of use; trust set up for a purpose	Membership	May enable consent; use of frameworks e.g. ADF (Anonymisation Decision-Making Framework)	Should focus on one or two classes of beneficiary

Article	Access	Purpose	Permission	Privacy	Value
Bunting and Lidsdell (2019)		Need for a body who decides who gets to use data and why			
Mulgan and Straub (2019)	Sharing, commercial, synthetic and research data; variety of access dependent upon use; hypothesis testing against data without access	Wide variety, no specific approach to deciding who gets to use it; households might pool data	Membership of data providers and users	Cryptography, digital twinning, technical and legal approaches as suggested by Young et al (2018)	Wide variety of personal, public and commercial applications
Stalla-Bourdillon et al (2020)	Use DPbD workflow: data minimisation, storage limitation, ensure processing is fair. Data pool	Comply with Article 5 (GDPR), define purpose, identify legal basis	Membership and (potentially) third parties	Organisational and technical measures as stated in Article 25 (GDPR), assess risk before processing, verify data processing	
EU COM 232 (2018)	B2G data sharing (public utilities etc). Data suppliers should support use to mitigate limitations	Limited to one or several purposes for limited duration; data will not be used for unrelated admin or judicial purposes, 'do no harm'	Contract		Clear and demonstrable public interest; mutually beneficial (compensation possible)
Fisher and Fortmann (2010)		Collective-choice arrangements (preventing harm to data supplier by poor use); overlapping claims can create bottleneck	Clearly defined boundaries with identification mechanism; sanctions; monitoring; licensing (through collective choice); affected by existing IPR and copyright issues. "Payment" in citation or co-authoring; subject to misuse		Scientific research - data collection and data analysis collaborations
Grossman et al (2017)	Datasets have digital IDs - these are associated with access controls; API access with authentication and authorization; portability between commons	-	Data Contributors Agreement and Data Access Agreement	May hold genomic and sensitive biomedical data - access after ethics committee approvals	Data intensive scientists - long term sustainability challenges; acknowledgement, scientific progress
Eschenfelder and Johnson (2014)	No sharealike, scholarly use only; registration required;	Some specific application for data use; repository and dataset level terms of use statements;	Community proxy makes access decisions; some institutional	Variation - some no sensitive data; some advice on PETs, some secure location for sensitive data; privacy	Some commercial use allowed as well as researchers

Article	Access	Purpose	Permission	Privacy	Value
	identified 9 different types of access/use rules	specific purposes (e.g. only diabetes research); some overlap with open	membership; original study manager may need to consent	was main rationale for control; respect of informed consent; confidentiality agreements	
Klievink, van der Voort and Veeneman (2018)	Parties exchange and integrate their data	Reporting requirements added by government (potential disincentive)	Moved to open data	Consent required; their case study platform moved to openness	New value beyond the immediate capabilities of the participating actors; public problem or demand
Schwabe (2019)	Blockchain consortia	Idea for a technological solution is the starting point (solution based probing)	Permissioned distributed ledgers; can automatically distribute tasks and support decision making on use		Similarity of members of consortia leads to similar value for all parties plus generic societal value;
Susha et al (2018)	Shared infrastructure; sets out a research agenda across these areas	Problem vs systemic?	Research agenda for this area	Anonymization, aggregation	Five value contributions: situational awareness/response; public service design/delivery; forecasting/prediction; evaluation/impact of policies; knowledge creation/ transfer
Susha, Janssen and Verhulst (2017)	Varying degrees of openness, selected users/processed insights	Specified/unspecified; on demand, event based or continuous	Agreement, application, open	Data about natural persons; potentially volunteered or user-generated	Realizing public benefit, rather than commercial innovation, value to data user
Verhulst and Sangkoya (2014)					
Susha, Janssen and Verhulst (2017a)	Must ensure useability, manage transfer and have method for resolving conflict	'Take it or leave it', mutual adjustment	Agreement, restricted environment, data stewards (trusted third parties)	Anonymisation	Mutually beneficial value
Duch-Brown, Martens and Mueller-Langer(2017)	Interoperability reduces market power	Negotiation	Contract	Anonymisation	Both parties must receive value or market fails
Richter and Slowinski (2019)	Third party vs company owned; multiple technical solutions; depend on	Inter-sector data marketplace;	Contract, financial transactions inc flat free, auction, regulation-enforced licensing		Innovation

Article	Access	Purpose	Permission	Privacy	Value
	interoperability, pooled; non-data holders can access pool				
Roman and Gatti (2016)	Parties jointly store and run computation on data while it remains completely private	Specific credit-checking sector and purpose	Decentralised and cryptographic technologies allow control and guarantees on use	Data protection by design and default	Speed up progression, innovative services, or payment
Fricker and Maksimov (2017)	Various	Various	Various pricing strategies		Profit (on both sides)
Stahl et al (2017)	Hosted by marketplace, who provide infrastructure; increase in web exchange formats		Free, freemium, pay per use and flat rate		
Koutroumpis, Leiponen and Thomas (2017)	One-to-one, many-to-one, one-to-many, many-to-many	Sector-specific address shared risk; admissible utilisation definitions hinder merging	Bilateral negotiated, 'bartered', standardized. Contract terms are contextual (based in specific jurisdiction). Rigorous provenance is proxy for control.	Sellers may not be aware of legal status of data vis-a-vis privacy	Enhanced market efficiency, resource allocation efficiency
Stahl, Schomm, Vossen and Vomfell (2016)	infrastructure that allows customers to upload, sell, browse, download, and buy	-	-		Note that few data marketplaces succeed
Carnelley et al (2017)	(e.g.) API access to linked, originally open, datasets	(e.g.) Agree to processing and use of data	(e.g.) Both parties must be registered and comply with marketplace terms of use		(e.g.) Smaller companies selling data for revenue streams, larger companies seeking insight
Schomm, Stahl and Vossen, (2013)	Multiple methods of access are key feature				

9.4.1 Access

In terms of how data is accessed, there is a variety of approaches deemed appropriate for use with trusts. For instance, Stalla-Bourdillon, Wintour and Carmichael (2019) discuss the use of data foundations with pooled data, while Stalla-Bourdillon et al (2020) discuss the use of data trusts with data that is entirely shared between holder/provider and user (as with the cities). Stalla-Bourdillon et al (2020) take the approach of being guided specifically by GDPR in the establishment of a data trust, so also note detailed aspects of data sharing that would ensure compliance with the principles of GDPR, such as data minimisation and storage limitation (indicating the length of time for which the data can be shared).

Data collaboratives take many forms, with the main definition being that they are a way for civic or governmental groups to access private data (Susha, Janssen and Verhulst, 2017). Consequently, data exchange and integration is an important feature.

How this is effected has many potential solutions. Stahl et al (2017) discuss third party hosting, (Richter and Slowinski (2019) data pooling, and therefore, interoperability is important). Providing linked data is also important because this adds value to previously open datasets (Carnelley et al, 2017). However, interoperability reduces market power (Duch-Brown, Martens and Mueller-Langer, 2017). The only paper which suggests data should be kept separately and computation run over it instead is that of Roman and Gatti (2016).

9.4.2 Purpose

Most authors specify that the data trust requires an explicit purpose for the data sharing. Stalla-Bourdillon et al (2020), guided as they are by data protection compliance, recommend identifying the legal basis for sharing. Bunting and Lidsdell (2019) focus on the process by which decisions are made as to what the purpose is and who gets access to the data. Data trusts are particularly concerned with third party stakeholders (authors often make the implicit assumption these will be involved in decisions) and with rights over the data belonging to individuals.

Eschenfelder and Johnson (2014) find that the repository mission is important in deciding access and use. Some may have very specific purposes - for instance, diabetes research - which immediately defines those who can access the commons as being those who are researching diabetes or associated issues. There is a similarity to data trusts in the use of collective choice, whereby those affected by the institutional rules can have a say in the operation of those rules, or may be represented by a community proxy to make access decisions (Eschenfelder and Johnson, 2014; Grossman et al, 2017). Eschenfelder and Johson (2014) also note that previous research showed scientists would be more likely to share if they knew who was using their data and could put conditions on access, reflecting that social interaction is important in data commons.

Susha, Janssen and Verhulst (2017) differentiate between sharing that is on demand (highly specified), event based, or continuous (less specified.) In later work this is characterised as 'problem versus systemic' (Susha at al, 2018). Klievink, van der Voort and Veeneman (2018) note the paradox of control and generativity - ensuring data is not misused also prevents some unforeseen positive use. However, they assert reconciliation can be possible through the flexibility of the collaboration, rather than the data. Schwabe (2019) suggests that the idea for the technological solution is the starting point rather than the data (solution based probing). While this makes sense, it can also be problematic. Susha, Janssen and Verhulst (2017a) offer the example of the partnership between Uber and the City of Boston, which was compromised due to the fact that the data shared did not correspond to the needs of the city. They suggest there are two fundamental approaches to sharing data - 'Take it or leave it,' (which is the default position of open data) or mutual adjustment.

Sector specific marketplaces address purpose and shared risk (Roman and Gatti, 2016; Koutroumpis, Leiponen and Thomas, 2017). In more general marketplaces, once negotiation of purpose is involved, attached admissible use definitions hinder merging with other data (Koutroumpis, Leiponen and Thomas, 2017).

9.4.3 Permissions

In terms of permissions, a membership model is most frequently cited by data trust researchers, although Young et al (2019) refer to a spectrum of permissions from ad hoc negotiations to membership. Hall and Pesenti (2017) assert that data trusts must reduce the transaction costs of accessing data, and a one-time negotiation of terms to access multiple datasets fits this requirement. They propound that this is necessary to ensure small companies can access data as easily as large ones, who are more able to absorb or circumvent access costs.

Data collaboratives are mostly seen as working through agreements (Susha, Janssen and Verhulst, 2017). However, data collaboratives work pays particular attention to the possibility of incomplete contracts (the unforeseen possibilities acknowledged above, especially in innovation) and ways of solving conflict (Schwabe, 2019). This requires some nature of organisational arrangement to address this. At the centre of this arrangement are data stewards, trusted third parties, analogous to but with more extensive responsibilities than the 'trusted GDPR advisor' of Hall and Pesenti, (2017) (Susha et al 2018).

In a marketplace the permission mechanism is largely contractual, usually for financial recompense. Some data is also bartered (usually in a many-to-one market, where data is traded in return for some kind of service. The contracts can vary from bilaterally negotiated or standardized. Koutroumpis, Leiponen and Thomas (2017) note that contract terms are contextual

- that is, based on a specific jurisdiction. Richter and Slowinski (2019) propose regulation-enforced licensing for certain societally beneficial uses.

9.4.4 Privacy

There is more agreement regarding the privacy aspects of trusts. All the publications agree that trusts are likely to - or are designed to - handle personal data, and they note a variety of technical and legal approaches. Stalla-Bourdillon et al (2020) state that technical measures must be combined with organisational measures to ensure compliance with GDPR, while Hall and Pesenti (2017) suggest that the data trust can act as a "trusted advisor" on GDPR matters. Stalla-Bourdillon, Wintour and Carmichael (2019) note that personal data and non-personal data are not binary concepts; non-personal data can become personal in a variety of ways, as described in Chapter 2.

In terms of privacy, scientific data commons may likely hold highly sensitive personal biomedical or genomic data. However, scientific data commons also benefit from academia's existing, well-established ethical guidelines for the collection, storage and sharing of data of even high sensitivity. The data is unlikely to be held without associated and very specific consents as to the purpose for which it can be shared. However, Grossman et al (2017) also assert that portability between commons is important.

Susha et al (2018) and Susha, Janssen and Verhulst (2017a) envisage uses for data collaboratives where technical solutions of anonymisation and aggregation would be provided and would suffice. Klievink, van der Voort and Veeneman (2018) suggest that volunteered or user generated personal data would be used, but would have appropriate consents attached.

Like Stalla-Bourdillon et al (2020), Roman and Gatti (2016) suggest data protection by design and default for data marketplaces. Their work focuses on alternative credit referencing data, so not only covers explicitly personal data but also the use of non-personal data for the creation of personal data. As noted above, they protect this by running computation over data without moving it. However, they are outlining a sector-specific data marketplace. In more general marketplaces, the privacy implications of trading data potentially limit their development, as ambiguities in the scope of the definition of personal data, coupled with an absence of personal data ownership rights or oversight, may create uncertainty. (Koutroumpis, Leiponen and Thomas, 2017; Duch-Brown, Martens and Mueller-Langer, 2017).

Young et al (2018) note that privacy standards are different in the rest of the world to Europe, and also differ between states in the US and this is important to bear in mind when attempting to draw conclusions about views on privacy between the different models of sharing, as the various rules prevalent in each area may affect the approach, even if the research focuses on multinational sharing models. As can be seen in Table 17, the data trusts authors are UK-based.

The data marketplaces work is European-based; the data collaboratives work is transatlantic and the data commons research reviewed is entirely US based. These geographic clusterings will be influenced by differing cultural, legal and political drivers.

9.4.5 Value

Like access, data trusts address a variety of beneficiaries, from innovation and economic (Stalla-Bourdillon, Wintour and Carmichael, 2019) to social (Hardinges and Wells, 2019, Mulgan and Straub, 2019, Young et al, 2019) and from many (Mulgan and Straub, 2019, Young et al, 2019) to few (O'Hara, 2019). The key message is explicitly that the data sharing must be "*mutually beneficial"* (Hall and Pesenti, 2017). "*Another motive behind the emergence of data trusts is the need to distribute the benefits from data use more equitably,*" (Hardinge and Wells, 2019). In other words, if value is only accruing to beneficiaries (the ones who are granted access) then the data sharing is not working. They point out that indirect benefits are impossible to redistribute, as noted with the employment and tax examples in the previous chapter. Therefore, more direct benefits (or value) must arise for both parties.

Value for data commons, again, is affected by existing rights and practices that pertain in academia. "Payment" for data might be made in citations or co-authoring. There is potentially quite a cost to sharing data in commons. If one individual or consortium uses the data to publish a set of results they can have copyright (or first publishing rights) over them.

Susha et al (2018) propound five value contributions from data collaboratives: situational awareness/response; public service design/delivery; forecasting/prediction; evaluation/impact of policies and knowledge creation/ transfer. The value must be in solving a public problem or meeting a demand that is beyond the capabilities of the participating actors individually (Klievink, van der Voort and Veeneman, 2017). Schwabe (2018) claims that the similarity of members of consortia leads to similar value for all parties plus generic societal value. Susha et al (2018) noted that only examining value for one side of the data sharing arrangement was a limitation of their early work. In this paper, they note it is *"counter-intuitive*" to expect private companies to consider sharing a worthwhile activity if there is only value to the data user, as noted in their taxonomy (Susha, Janssen and Verhulst, 2017a).

Like any market, most authors find the transactions are largely driven by profit and efficiency. Carnelley et al (2017) note that smaller companies are in these markets in order to obtain direct revenue while larger ones seek insight to enhance or create new offerings. The innovation potential offered by data coming to the market in this way is an important feature of marketplaces (Roman and Gatti, 2016, Richter and Slowinski, 2019). Stahl et al (2016) note that few succeed.

One last report should be considered here, as it emerged in the citations but is neither trust, collaborative, commons nor marketplace, yet bears similarities to them all. EU COM 232 (2018) refers to 'data spaces' in Europe. Like collaboratives and trusts they focus on mutual benefit, like marketplaces they consider the possibility of compensation. Similarly to trusts and collaboratives, purpose is limited in time and scope. Like collaboratives, they seek to promote business to government data sharing.

Additional to the concepts identified in the matrix, there were strong governance and use themes that appeared integral to each format. The data commons work had a particular focus on collocated tools; data collaboratives on data stewards, data trusts on the involvement of stakeholders in decision processes, and data marketplaces on monitoring.

9.5 Summary

This chapter describes the findings of the integrative literature review of the data sharing literature, following the methodology outlined in Chapter 5. Using a concept matrix based on the framework of open data for open innovation presented in Chapter 4, it shows an analysis of the literature identifying themes; patterns, such as the consistent emphasis on two-way value and relationships; contrasts, such as the variety of ways of dealing with personal data, and evolving the particular from the general.

In the next chapter, I discuss these findings to produce a conceptual framework that creates insights for filling the spaces left by open data in public sector open innovation.

Chapter 10 Discussion - Does comparison of other types of public and private data sharing arrangements enable revision of the previously defined theoretical framework of open data?

While there are a number of existing models - and potentially, another in the shape of 'data spaces' – these exhibit both overlaps, such as a broad view towards modes of access and a range of permissions, and specific differences - generally around value and privacy. Below I discuss what insights data sharing can offer for open data.

10.1 Personal Data and Data Sharing

One area in which open data appears particularly challenged is that of personal data. As shown in Chapter 6 and also in the work of van Loenen and Welle Donker (2012), data holders frequently do not know how to deal with sensitive data. Reidentification presents challenges to anonymisation, and non-personal data can become personal via use.

Data that is going to affect humans in any but the most tangential way is going to be, at least in Europe, personal in some way, and this is likely to increase with utility. As an example, train arrival times are not personal data if they are read from a web page. Train arrival times that are being digitally delivered to someone for a specific reason constitute personal data. Sensor data will further complicate this. It may be that the vast majority of generally applicable, useful, frictionless data has already been released. The possibility of successfully anonymising/aggregating personal data while still retaining usefulness is largely assumed by most papers in the review. Stalla-Bourdillon et al (2020) provides most detail on how deal with actual personal data that arises, via a 'Data Protection by Design' (DPbD) approach, building in consent, data minimisation and storage limitation.

The papers included in this review largely did not specifically talk about sensor data. However, as the volume of sensor data grows, two things might happen. Firstly, the relevance of non-sensor - static, historical, frequently exhaust data (the kind that is often opened) - may decrease. Secondly, understanding of what type of sensor data should be counted as personal, will increase.

However, data-related risks extend much further than privacy alone, for instance, intellectual property rights clearance and management, anti-competitive practices, contractual compliance and confidential data management (Stalla-Bourdillon; Wintour and Carmichael, 2019; Koutroumpis, Leiponen and Thomas, 2017). These are seen specifically in data commons, which betrays its name by its struggles with zero sum games (i.e., data is in no way reduced by use by

another party, but publishing first with the data might increase gains), but also in City 4, where there were issues with the citizen report usage.

Data accompanied by pre-agreed use therefore has limits on use. Without pre-agreed use, there is a limitation on what data can be shared. As with Susha, Janssen and Verhulst's (2017a) Boston and Uber example, data that is available is unlikely to be a perfect match if the use comes first. However, use first is likely to be the more valuable opportunity. Therefore, speeding up the availability of data, once that data is decided upon, is important. One part of this is a smooth mechanism for deciding open innovation use, whether that is a one-to-one situation, a data trust or other situation with enough flexibility to make informed decisions, which will almost certainly involve humans either setting up the rules that can then be applied in smart contracts or making the actual adjudications.

Such a mechanism for sharing should be based around mechanisms for developing the relationship, such as inbound and coupled open innovation. Parties with data to open should consider to whom they wish to open it and why, in order to incentivise internal cooperation with opening, and to create the highest likelihood of use. Any data that can be made available for anyone to use for any purpose can subsequently be opened - having first been tested. Thus, the 'purpose' element of the framework becomes part of the trajectory of open data. Time limitations and storage rules should be imposed in case it is necessary to 'unshare' or stop the opening of data.

Young et al (2018) repeat the critique of open data by some open government advocates 'self-selecting' and being unlikely to disclose certain types of information, much as the cities were averse to releasing the air quality data. Encouraging data holders to begin to collect and generate data in a way that will be easy for them to share technically and legally, should the economic or social incentive arise, is the preferred approach. Identifying the legal basis for sharing (as recommended by Stalla-Bourdillon et al, 2020) has much in common with the process of opening government data, which is argued should be open as it is generated by tax payer funds.

10.2 Institutions and Governance

Mulgan and Straub (2019) talk about "public interest institutions". However, further physical/digital institutions may reduce the ability of cities to develop flexibly. Preparing their data to be pre-deposited in a trust where a third party makes decisions about who gets access will lead to similar problems as exist with open data – it is too much work, it is too generic, they do not necessarily know what to prepare. Nor does this address the issue of when cities realise that a third party, and not they, hold the data that is needed. They need a data sharing method that can be responsive to need (and that will give any third party confidence to also share their data).

Moving to a framework where use and access are initiated together - at least initially, and potentially always - will reduce up front work that may be perceived by some of the city employees as wasted.

Data trusts and collaboratives take a third party, stewardship approach, which may be applicable here. Data marketplaces act as technical and legal intermediaries that may also have a use (although Dawex sees itself as an 'AirBnB of data'). None of these formats suggest direct data sharing. While this is frequently done (for instance, in the Horizon 2020 Data Pitch project) it does not scale well, and leaves scope for error.

Post-availability data governance is key. Stalla-Bourdillon, Wintour and Carmichael (2019) note that the release-and-forget model (of open data) is not likely to be conducive to best practice for data governance. For instance, it will be difficult for data providers to ensure the data user complies with reusage restrictions, and that anonymised data are not re-identified. Koutroumpis, Leiponen and Thomas (2017) note that as the value of data critically depends on the appropriateness of the procedures associated with collecting, organizing, and curating it, knowledge about the origins of data is usually critical to discern their quality and protection status. In open data, where datasets can be (and often are) rehosted, reversioned and generally replicated, the single version of truth (and associated documentation, if it was ever there) becomes ever more elusive. Google Dataset Search, launched at the end of January 2020, aims to improve the ability to find original datasets, but shared data only amenable to approved replication, and therefore it is theoretically more likely to retain a single version of truth.

Mulgan and Straub (2019) argue that that we need, but currently lack, institutions that are good at thinking through, discussing, and explaining the often complex trade-offs that need to be made about data. Koutroumpis, Leiponen and Thomas (2017) suggest that in the absence of other controls collective multilateral governance forms might be expected to appear. Stalla-Bourdillon, Wintour and Carmichael (2019) write that the data governance model must have a decision-making body that engages participants (in particular data providers) and represents the interests of data subjects. In a public sector scenario, this might formally include citizens, or it might simply look as it did in SCIFI, where citizens were consulted about the challenges, and the selection of the data user was left to the city representatives.

10.3 Additional Frictions from Data Sharing

Two major frictions that open data was supposed to ease were license combination and discovery. So far it has been shown that license combination continues to be an issue as (a). there are multiple open data licenses and rights, and (b). that a closer relationship with either the data holder or their intermediary is valuable for successful use, whether this is because of necessary

domain or documentation knowledge, or because it is not made open until the relationship of use is established.

Open data has the potential to reduce excessive first mover advantage. In the absence of open data, large companies are often better placed to achieve this advantage, 'Towards a Common European Data Space' (2018) aims to require a more transparent process for the establishment of public-private arrangements, to align with this. The main problem of such arrangements is that in practice they lead to one, or very few, reusers exploiting the data in practice and that this limited re-use is due not to the specifics of the market but to the way in which the public- private arrangement was concluded. While this it true - and City 4 queried how sharing data might cause failure in the competitive market - the public-private arrangements can be altered without data pooling.

Data sharing does introduce some new frictions. Fisher and Fortmann (2010) point out that overlapping claims to data use can create bottlenecks. Klievink, van der Voort and Veeneman (2018) note that a multiplicity of actors and goals increases the likelihood of conflict between those goals. Young et al (2019) note that contract terms are contextual - that is, based on a specific jurisdiction. This could be even more complicated than varied licenses, but a third party or data steward in a particular jurisdiction which pertained over the entire data sharing collaborative, or trust, or other model, could eradicate this. Another friction suggested by Koutroumpis, Leiponen and Thomas is the cost of monitoring. However, this too can be managed, in the context of the ongoing data governance which is required.

10.4 Data Sharing for Value

In some ways, the data sharing models are a way to focus a data ecosystem for use. In the open data ecosystem, it is not just the data that is important, but the processes around it. In the 'Open Data Incubator for Europe Impact Assessment' report, the second most valuable service provided to incubated companies, after funding, was "*access to the open data industry*" (IDC, 2017). The data sharing models clarify that direct value has to occur for both parties. This is particularly important where non-tax payer funded (i.e. commercial) data is shared, but also because much local government and agency data has a very long and circuitous route to impact in terms of benefiting the holder from tax or employment. Consequently, the mechanism for sharing has an important role to play, as it both sets up the value benefit dynamic, and ensures data sharing does not simply become a request system with some form of decision and tracking. It further requires some kind of direct payback to justify the cost. However, this does not have to be a formal institution, as noted above. Verhulst and Sangkoya (2014) identify the open innovation mechanism of digital innovation contests, in their words, "*prizes and challenges"* as a type of data collaborative. Within this context, a particular workflow - a series of accountable decisions and

actions taken by responsible individuals with appropriate expertise prior to the commencement of the data processing activities must be in place (Stalla-Bourdillon et al, 2020).

10.5 Conflicts with the Original Framework of Open Data

Considering these adaptions suggested to the open data framework, it is important to ask how far they are antithetical to the characterisation of open data. Longshore Smith and Seward (2017) characterise openness as 'you don't have to pay' and 'anyone can participate'. None of the above models require payment (although as previously noted, there should be a clear path to return on investment). The second aspect is definitively a theoretical statement, as there are already limits on 'anyone can participate', and there are always rules of access that have to be adhered to. At the least, the 'price of entry' is having the data literacy skills to understand data driven innovation, manipulate datasets and use the various necessary tools (Argast and Zvintsayeva, 2016; Frank and Walker, 2016). Similarly, in open innovation, there is often a wide net for participation at the proposal stage, but actual engagement is defined by the funnel process – only those that achieve certain stages participate.

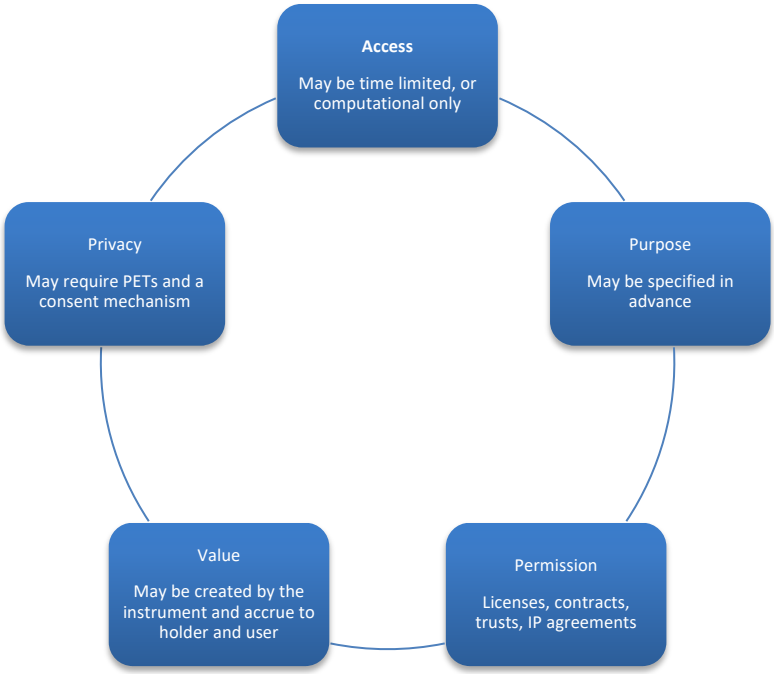

Figure 18 Adapted Framework of Open Data for Open Innovation

These considerations lead to an adaptation of the open data framework as shown below.

Table 45 Original and Adapted Framework of Open Data for Open Innovation

Concept	Original Framework	Adapted Framework
Purpose How purpose is decided and monitored (this will be affected by value mechanism).	Open data is agnostic as to use. This is perceived as a strength for innovation, as it does not curtail any use nor focus invention in a specific area.	Purpose may be decided and monitored (this will be affected by value mechanism). The purpose of the use of data must be compliant with legal basis for processing data and take into account the creation of personal data. Other security issues must also be considered.
Access How the data is physically accessed and shared	Access is equally available to all. Provided people are skilled enough to discover the data and then utilise it, they can use it. Multiple programmes exist to support this. Again, this is seen as a positive aspect for innovation, as it removes first mover advantage that might accrue to larger firms with greater resources.	It may be that the data is sensitive and may only be accessed through specific platforms, or only have computation run over it. The time period for sharing must be defined, and the end of the period managed – whether it is closed or subsequently opened. Reducing the movement of data must also be considered.
Permission	Although mention of a license is not explicit in the open definition, data without an open license confirmation is simply publicly available information, so this is a core requirement. Again, this is a strength for innovation as often, (although as seen in Chapter 2, not always) there are no rights-based complexities associated with this.	This may be a license that is granted (including open), a legal trust, a contract or an intellectual property agreement.
Privacy	It is intrinsic to establishing a basis of comfort with the idea of public sector information being made available for use (including innovation) that open data is not personal data. Privacy is therefore a boundary of open data - once this is crossed, it cannot be open data.	Compliance with GDPR in terms of pseudonymisation/anonymization where necessary. Must comply with the purposes consented when the data was collected. Requires a mechanism enabling this to be confirmed and approved.
Value Instrument		Value lies in the innovation mechanism for which the data sharing is established. This

Concept	Original Framework	Adapted Framework
		must be compliant with permissions and purpose.
Data Holder Value		Value that accrues to the data holder.
Data User Value	While value is not intrinsic to open data itself, it is the key motivator, for opening and use. The complexity is in defining where the value lies. Often the value is indirect (largely the value is indirect through tax payer benefits), or, per Lassinantti (2014), widely distributed. It is often difficult to identify where the locus of the value is in open innovation with open data because unless a coupled open innovation mechanism is used, tracking is not possible, and there are few impact models.	Value derived by the data user.

The above is a suggested framework for (open) data sharing, based on praxis, that moves beyond a privileging of open data and integrates it with other models of data sharing. Based as it is in public sector open innovation, it requires resilience testing against other forms of data-driven innovation undertakings. This has been done for certain time-limited, government-propelled initiatives (Walker, Simperl and Carr, 2019). However, there are entities which control and produce vast amounts of data, such as Google, Facebook and other social media companies, whose practices have not been considered here.

10.6 Summary

In this chapter, I discussed how data sharing can support the use of personal data, and also its limitations, as found in the literature, for dealing with sensor data. Conflicts of data sharing with open data were reviewed, and found to be more theoretical than practical. The role of data sharing for value was examined, in particular through open innovation, and it was established that clear and mutual benefit is necessary. Additional frictions may be created especially around multiple users. Institutions and governance play an important role. Finally, I presented my key findings and an updated framework of open data for open innovation.

Key findings of this research:

1. Some authors (and practitioners) already understand open data as a form of data sharing, which has specific limitations and permissions vis-a-vis privacy and licenses, and no limitations vis-a-vis access and purpose. This approach does two things. Firstly, it removes the privileging of open data as a special institution, an activity which is somehow separate from other data sharing activities. Secondly, it offers a spectrum along which the realities of open data can be placed.
2. There is no way of completely reducing friction. It is not possible to just 'release and forget' (Stalla-Bourdillon, Wintour and Carmichael, 2019). This is true even for open data, which, in common with other data sharing forms, requires a managed ecosystem of users around it, engaging in the "mutual adjustment" identified by Susha, Janssen and Verhuulst (2017a).
3. Once this managed ecosystem becomes more formal, i.e. once there is a system for selecting who should get access, and there are mechanisms for managing that access from a legal and technical point of view, governance becomes a key issue. Data Protection by Design assists with this.
4. Mutually beneficial value models are the key for data to be shared and used. It is not incumbent on data sharers and users to be able to necessarily identify that value. Third parties who act as promoters and intermediaries may take this role.
5. Certain kinds of value can only be derived from certain mechanisms. Certain benefits will only align with certain forms of data sharing. Broadly, in a marketplace, few organisations will be incentivised by a co-authorship on a paper; conversely, in a commons, it would be extraordinarily unethical to vend patient data. More narrowly, data sharing for open innovation via any mechanism requires a specific set of potentially enactable benefits for both parties. Therefore, the value instrument matters for establishing who benefits from what value.
6. Given the above, the relationship between the data sharer and the data user must be at least as important as the data itself.
7. The third parties who act as promoters of data sharing mechanisms or intermediaries may be data stewards or act as data/GDPR advisors, as suggested in work on both collaboratives and trusts (Susha, Janssen and Verhuulst, 2017a; Hall and Pesenti, 2017).

Given issues of sensor data, the current scenario most likely represents the beginning of the privacy/personal data journey for organisations, at least in Europe. The next chapter presents my conclusions and future work

Chapter 11 Conclusion

11.1 Aim of Thesis

The general objective of this thesis was to understand how the processes of the use of open data for open innovation are operating in reality and locates the conditions that may cause it to diverge from the use presented in the literature. It investigates this space, and derives a framework of open data that reflects productive yet legitimate use.

The specific research questions were as follows:

RQ1: What are the key components of a framework of open data for open innovation?

RQ2: How does the use of open data in open innovation in practice vary from the framework defined in RQ1?

RQ3: How can comparison of other types of public and private data sharing arrangements inform the framework defined in RQ 1 so it more accurately reflects open data for open innovation as found in practice?

11.2 Review of Research

In Chapter 1, I set the context for this thesis in the relationship between open data, open innovation and the public sector, and provided evidence from the literature that the uses and impacts had not transpired as anticipated over the past decade of open data.

Chapter 2 offered a detailed review of the open data literature following the process from the motivation to open through publishing to use. In Chapter 3 an overview of a specific form of innovation, open innovation, was presented, with a particular focus on public sector innovation and smart cities. Chapter 4 synthesised these concepts, using the emergent literature in this space and examples, to develop a framework derived from the previous literature chapters.

Chapter 5 presented the background to the case study utilised in the original research. In Chapter 6 the methodology to be followed to answer Research Questions 2 and 3 was outlined. RQ2 was addressed with a case study of public sector open innovation in 4 European mid-sized cities which were part of a consortium, with the aim of increasing the framework conditions for smart city focused open innovation. RQ3 was addressed with an integrative literature view of data sharing literature. The benefits and limitations of both approaches were outlined.

Chapter 7 presented the findings of the metadata analysis, the document analysis and the group interview that comprised the methods to investigate Research Question 2. It used the metadata

analysis to inform the document analysis, and then used the group interview to comment on and confirm the combined analysis.

Chapter 8 asserted that for the municipalities, it is not realistic to open any dataset for any purpose. There is too great a legal, technical and administrative work load required, and insufficient reason for co-operation from colleagues who have competing demands on their time. Opening by default may lead to political risk; the incentive is at best vague and it may lead to privacy risk. In the cities this led to either paralysis, with no data sets being opened; with few, limited and 'safe' datasets being opened, or datasets being opened for a specific purpose in combination with stakeholders and users. Even mandating certain datasets to be opened at the municipal level is difficult to monitor for compliance.

In Chapters 9 and 10 a framework of open data that addresses the challenges of the GDPR, new technologies, licenses and benefits was devised and discussed. This was done via an integrative review of the data sharing literature, focused around the original open data framework established in Chapter 2. It concluded with an amended framework of open data for open innovation.

11.3 Implications

A question that was asked during the defence of this thesis was, "Are you suggesting the end of open data?" Given that the amended framework essentially borrows approaches from data sharing, rather than open data, and that the overall approach abandons the quest to improve the open data publishing and use process in order to embark on a quest to understand how the somewhat ad hoc process that is being undertaken can be legitimised, this is a crucial comment.

My research is inspired by two quotes from the literature. The first of these is, *"Open data is more complicated than presented,"* Davies (2010). The initial framework presented in Chapter 4 underlines the truth of this – it is comparatively easy to strip the concepts of open data back to their bare bones and present it in a simple fashion, where privacy is never compromised, licenses seamlessly interweave and everyone has the skills and opportunity to access the data that they need for a variety of valuable purposes. I stated in Chapter 1 that the two main constraints on this are resource and regulation.

In the pilots researched in this thesis, the data is often not open at the beginning of the projects because of a lack of knowledge about what to open, and often not open by the end of the pilot because the process takes too long.

Resource constraint is a fact of life, but at the same time, resources can always be found if the matter in hand is sufficiently important to the data owner. In Chapter 1, I referred to Janssen, Charalabidis and Zuiderwijk (2012)'s critique of the 'if we build it, they will come' fallacy. Simply

opening the data is not enough – it is necessary to add a good ecosystem of local (and not so local – one of the successful Dutch pilots was run by an Italian company) SMEs, and to contribute city knowledge, time, and support. Even the relatively minor resources of SCIFI have enabled 14 pilots to go ahead, of which 4 solutions are currently in the process of being procured by the cities. Some of the data in these projects was not open at the beginning of the pilots, and some was still not open by the end, because of the length of time the process requires. However, the success of the pilot makes it more, not less important, for the data to be opened. With shared data the city is just tied to one provider, and is a potential victim to market failures (such as a SME going bankrupt due to Covid). Resources are therefore found to ensure the data is opened, and open innovation incentivises the opening of the data, becoming a virtuous circle.

Regulatory constraint, and worse, constraint caused by confusion over regulations and matters of risk, liability and exposure, is much more problematic, as it is much more likely to prevent the pilot from successfully launching. While plenty of existing open data, especially geodata, is unlikely to be personal, there are three major pain points where personal data may occur: sensor data, data combinations and data reuse. In terms of the second two points, these risk existing open data being transformed into personal data. If this happens, there is no protection from liability for the user. In a case too late for inclusion in this research, a city decided not to use a third party's data because they felt exposed to such risks, despite assurances that it was safe, compliant and non-personal. This is why, despite their being a tool of data protection, the SCIFI cities utilised data inventories, to understand what problems might be caused by theoretically non-problematic data.

As there is no way to guarantee that open data will never be personal data, we may have to build data protection by design into our technical and organisational processes around open data, even though such processes are generally considered in relation to data sharing.

Thus, the second quote that is relevant here is, *"Open data is not a binary, but a range,"* (Eaves, 2013). As shown in the previous chapter, it can be seen as an entirely artificial construct that we privilege open data and separate it from other modes of data sharing, and open data can, in fact, be understood as occupying one end of a spectrum of sharing. Once it is on that spectrum, open data is not limited to one place. By being aware in advance that open data might appear in many guises, perhaps with compulsory registrations, or a combination of licensing, or with limitations on what it can be used for to avoid becoming personal data, preparation can be made for the inevitable frictions. While it would be clearly ideal that each additional set of open data that was being used came with identical licenses and no privacy risks, it is far more pragmatic to prepare systems and processes to deal with variety and edge cases, rather than to experience this as a problem.

11.4 Relevance

The case study of the Smart City Innovation Framework Implementation and the research within this thesis is situated between an established area of investigation of publication on the one hand, and reuse on the other.

On one side, scholars whose work addresses improving the publication and use of open data by identifying and removing barriers will recognise some of the issues contained in my research. Scholars who locate the value and impetus in open data firmly within use and look at improving strategies for reuse to do this will also find some familiar issues.

However, both these perspectives work within the proposition that the onus is on publishers and users of open data to achieve the model of open data, which is, by implication, functional and appropriate. This thesis takes a less normative approach, by investigating the publishing and use of open data in a specific context - that of civic open innovation - and asking how open data is working in practice. It then seeks to discover what praxis can reveal about open data use. It finds that while people are still very knowledgeable about the Platonic ideal of open data, in practice, they find other ways to achieve their open data goals and receive the benefits. It makes the key assertion that these practices are highly identifiable, and that they can be identified in the data sharing literature, which aids in amending the open data framework.

It sets this all in the context of open innovation, of built-in publication and reuse, while seeking to be positive rather than normative, reflecting reality rather than proposing how open data publishers should act to be more effective within the existing framework. The main contributions of this research are firstly that it seeks to combine open data and open innovation in the literature; it adds to the literature on public sector open innovation; it proposes a new lens with which to view the lack of impact of open data; and importantly, proposes concrete suggestions for enacting changes to open data structures that will enable greater productivity of data-driven innovation.

11.5 Limitations and Future Work

The main limitation on this work is that it uses a public sector open innovation case study. The framework for companies that hold a vast amount of data, but which have not published a great deal of it, such as Google and Facebook, may well be different.

As shown in Chapter 2, up to 7 different kinds of benefits of open data have been identified in the literature. Open innovation in the public sector is a subset of both improved/more efficient public services and innovation/economic growth. However, one of the central uses of open data –

perhaps the main use globally – is that of transparency. It is important to acknowledge that many of the suggestions proposed above could be construed as problematic for transparency.

Open data is beneficial for transparency as it enables activists, non-governmental organisations and charities to obtain information about spending, corruption and law-breaking activities across all levels of government. Being able to obtain this anonymously protects activists from reprisals and prevents governments from refusing to share data with parties they do not approve, such as unsympathetic journalists, hence open data is a useful tool of transparency in its current form.

The example of the air quality pilot makes an excellent use case for exploring this. As seen, City 1 does not have a well thought out strategy for addressing air quality issues in the municipality. They are thus unable to manage citizen's expectations regarding air quality and so are reluctant to open the data set. City 1, however, is in a country with a wide variety of governance and jurisdiction structures at the national and supranational levels. There is also (raw) open data on air quality made available from the European wide Copernicus open data programme. In other words, citizens are not having any information withheld from them, should they wish to seek it out. Equally legitimately, however, it can place unfair burdens on governments to open data to which they are not in a position to respond.

Future work could explore this tension between the aim to be 'open by default' and increase transparency, and to be 'open by experiment' and support innovation. However, it is largely likely to be a debate that plays out in philosophical or political arenas, as the same rules that prevent certain data from being opened for innovation will apply to data being opened for transparency. The question is whether there are some governments who will take advantage of this in a negative, rather than positive, sense.

Another vital area of future focus is working with sensor data. How can data holders assess which sensor data can be opened and which is too much at risk of triangulation? Is this possible? Will the dominance of sensor data fundamentally change the role of open data? It is difficult for cities to resist relying on (relatively) clean and structured sensor data versus messier historical data that may need its relevance teased out. This has the potential to hugely affect what cities are willing to open, as one said, "*sensor data versus existing data affects the focus*".

Chapter 2 of this thesis gave a brief overview of the open data business model literature. There is scope for merging the framework developed here with the existing business model literature to discover new opportunities.

Finally, this work reviewed the data sharing literature to reconceptualise a framework. It does not seek to promote any particular form of data sharing as a kind of solution to the challenges facing open data for open innovation. Data sharing, as it works today, is not perfect. How a suite of data solutions can sit side by side, be effectively communicated and be resourced, still requires investigation.

11.6 Academic Contributions

This thesis has made a number of contributions to both the open data and open innovation literature, and to that of data sharing.

As well as adding to the literature on public sector open innovation, it is one of the first pieces of scholarship to combine open data and open innovation in the literature, exploring the ways in which previous researchers have mapped open data activities to open innovation activities, bringing these together and exploring new ones.

In terms of open data, it advances a new lens with which to view the lack of impact of open data, which proposes that supply and demand need to work together simultaneously. In creating a new framework, it provokes theoretical debate, and by looking not directly at the barriers to open data publishing and usage, but by examining the routes people take to circumvent those barriers, it has initiated a new way of approaching open data research. It also provides the first integrated literature review of data sharing models.

Finally, it makes concrete suggestions for enacting changes to open data structures to enable greater productivity of data-driven innovation.

11.7 Policy Contributions

The cities were very aware of best practices of open data. They had issues complying with it, and little idea of the pitfalls where they diverged from it. They also struggled with some of the legislative aspects of open data, such as mandated data set opening, or streamlined licensing. Therefore, this suggests that horizontal (that is, non-sector or industry specific) non-legislative measures, such as awareness-raising and the sharing of data-driven innovation best practice, might be an appropriate policy solution.

In some cases, the cities did take the data that they had that they were not able to open, and set up appropriate contracts, intellectual property rights and confidentiality protection. Much of the time these were at least partially in place, but not all, especially around new things like sensor data. The cities also immediately stopped opening data where they realised it contained personal data, but it is also possible that they could have taken steps to, for instance, assess what consents were related to the data. This suggests the need for a kind of ‘virtual intermediary,’ which they could consult for the correct guidance. The increased availability of the artefacts of data sharing - contracts, DPbD guidance, data minimisation checklists and so on - would assist.

It is important to avoid the development of some form of arbitrary, opaque (and resource intensive) ‘request’ system for data to be shared or opened. This can be avoided with an open innovation programme mechanism that allows for use, opening, testing and monitoring while

being productive. This will generally be around a decision on what will be done with the data, and the agreement of some form of compliance. Where datasets could ultimately be destined for opening (e.g., they contain no personal information) this may facilitate speedier uptake. Where there is an established two-way relationship between the publisher and user, the user may be happy to accept rawer, less 'clean' data, thus moving those costs away from the publisher. It reduces the need for offering the data in multiple formats, and reduces the risk of re-identification inherent in anonymised open data. As the publisher is aware of the intended use of the data, tracking the impact will be simplified, thus addressing the measurement problem.

11.8 Summary

The praxis of productive open innovation with open data diverges from the theoretical approach. Opening data with a license and calling this outbound innovation may fit the model, but it is impractical in terms of the sustained creation of value. The GDPR and license variation are reducing the value of open data in order to retain regulatory compliance. The sometimes poor fit between existing data and valuable problems privileges the relatively new (and therefore, potentially problematic) sensor data. The apparent simplicity and reductionism of open data is actually messy in practice. Even when data publishers believe in a 'perfect' model of open data, they cannot dedicate themselves to this model if they want to achieve results.

Useful open data, therefore, requires a new model. One has been shown here that is derived from an analysis of the needs of data sharing that will assist with data being opened gradually, safely, purposefully and legally, creating a relevant data governance framework.

Does this, in some ways, depart from the 'spirit' of open data? It certainly conflicts with 'open by default' as outlined in the Open Data Charter. However, as this thesis shows, 'open by default' is a theoretical ambition, rather than a reality, for the cities in the case study. It could be argued that it, in fact, does not change the data that is eventually opened at all. It simply offers a way for more useful data to enter the data value chain, and promotes safeguards against the opening of data which should not be opened.

Appendix A Metadatasets Template

Identifier	**Data Access**	**Format**
Uniquely identifies the dataset. Could be a URI, if not use the format Dx.y, with x referring to the number of the challenge, and y identifying the dataset within the challenge	What are the access rights to the data: is it public (so open data), restricted (can be used, but with certain restrictions), or non-public (closed)?	The machine readable format of the file: csv, json, kml,
Challenges	**Update Frequency**	**Size**
The Challenges for which the dataset could be useful.	The data periodicity: One-off (was measured only once), intermittent, at regular intervals or even in real-time.	The approximate number of records to be found in the dataset
Title	**Temporal Coverage**	**Attributes**
Concise description of dataset content	The timeperiod covered by the dataset	Which attributes can be found in the dataset?
Keyword	**Geographic Coverage**	**Privacy Concerns**
Single word describing dataset content	The geography covered by the dataset	Will any attributes need to be anonymized?
Description	**Granularity**	**Data Quality**

Remarks

Data Inventories

City 3 Data Inventory Call 1 Challenge BC5

A	B	C	D	E	F	G	H	I	J
	OPEN DATA PUBLICATION TRACKING								
Name of challeng	Name of data user	Dataset name (	Dataset name (FR)	Description of the content	Owne	OD	Pub	Publis hed	Comment on data
Watering optimizat		Events calendar related to occupation of public green	Calendrier d'occupation des terrains de sport	Weekly calendar concerning the occupation of field sport : name of the field, timing slot (start and ends), name of the activity	City	N	N		Not possible to publish this dataset because of security concerns according IT departmeent. So they remain closed data
Watering optimizat		Wether forecasts	Prévisions météo	Weather forecast for the next 3 hours	Meteo provider	Y	Y		Public data service managed by wether operator.
Watering optimizat		Humidity of soil	Humidité du sol	Sensor measuring the humidity of the soil	City	N	Y	SCIFI datahub	Data gathered by sensors deployed during the pilot. Data are published on the SCIFI data portal They remain closed data as this is just an
Watering optimizat		Ambient air temperature	Température de l'air ambiante	Sensor measuring the prevaling air temperature on the sport fiels	City	N	Y	SCIFI datahub	Data gathered by sensors deployed during the pilot. Data are published on the SCIFI data portal They remain closed data as this is just an experimentation
Watering optimizat		Electric voltage	Tension électrique	Measure of level of batteries of humidity sensors	City	N	Y	SCIFI datahub	Data gathered by sensors deployed during the pilot. Data are published on the SCIFI data portal . They remain closed data as this is just an experimentation

Subset of City 4 Data Inventory Call 1 Challenge BC7

piroutes									
Dataset name (ENG)	Dataset name (NL)	Request applicant in Milestone ?	Owner	Open data	Publisl	Published where (URL of URI)	Date of publication or sharing	Difficulty opening up the data?	Questions and remarks of Qua
Pathways	Paden en wegen	Y	City	Y	Y	BGT	publication	To be confirmed by Herman	Used for visualisation purposes in PoC
De-icing routes	Strooiroutes	Y	City	Y	Y	City	publication	Done	Used for visualisation purposes in PoC
Traffic	Verkeersintentsiteit	Y	City	tbc	tbc	Shared		Work in progress	Verkeerslussen?
Sensoring information from salt spreaders.	Sensorinformatie van de strooiwagens	Y	n.v.t.	N	N	N.v.t.		We do not have the data	
Weather data	Weerdata	Y	Meteoconsult	tbc	N	Via contactperson Meteoconsult		Work in progress	Contact has been made, data sharing hard; not open data
Weather data (2)	Weerdata (2)	Y	KNMI	Y	Y	KNMI	-	Done	
Infrared data of roads	Infrarood data of roads	N	City	Y	Y	$	Publication	Done	Used for visualisation purposes in PoC
countings of cyclists and cars	tellingendata van fietsers en auto's	N	City	Y	N	Shared		Metadata	
Smart traffic lights (talking traffic)?	slimme stoplichten tellen, fietsers en autos	N	City	N	N				
Countings of cars (inside the road)	Teldata van lussen in wegen	N	City	tbc	N	Shared dataset	Shared	Work in progress	Received, but never applied in actual model
Data of roadworks / local traffic control	Nederlands wegenbestand	N	City	Y	N			Work in progress	

Appendix C Topic Guide for Group Interview

Section	Content
Purpose	Confirmation of metadata analysis accuracy Views on metadata analysis Views on key issues that arose that diverge from open data practice
Timing	45 minutes – 1hour
Introduction	Welcome, explain format of session. Share participant information and consent forms.
Metadata Exhibits	Share metadata exhibits (ppt). Explain methodology. Gather responses.
Sensor Data	Share findings on sensor data (consent to use; access to; potential for being personal data). Gather views.
Personal Data	Share findings on personal data (sometimes exists within less sensitive datasets; potential for becoming personal data via use). Gather views.
Sharing Data	Share findings on data sharing (it seems easier than opening data; there is inconsistency over how this is managed in contracts). Gather views.
Closing Discussion	Is there anything I have not touched upon that you would like to add?

Appendix D Thematic Analysis Codebook

Node	Sub-code	Description
Access		
	Availability	Internal processes of locating/collating data to make it public
	Publishing Decisions	Deciding what to publish
	Discoverability	How open data is promoted to potential (external) users
	Standards	Standards and interoperability
	Risks of Access	Caution motivated by concerns other than personal data
	Quality	Internal assessments and understanding of quality
	Organisation	Internal processes affecting publication of open data
	Sources	Desirable data owned by third parties
Purpose		
	Guided Reuse	Promoting possible ways to reuse data to external audience
	Selected Reusers	Engaging specific reusers in reuse of data
	Tracking Reuse	Establishing what data has been used and what for
	Dogfooding	Using their own data internally
	Collecting/Generating New Data	Sensor data
Permissions		
	Licensing	Relating to use of licenses, types of licenses
	Uncertainty	Evidence of lack of clarity around open status of data
	Ownership	Comments about ownership of data that are not about licensing
	Shared data	Data that is not fully opened but is shared
	Elision	Instances of open data being conceptualized as a form of data sharing
	Unacknowledged sharing	Where data is intended to be open, but does not meet the full criteria
	Benefits of sharing	Comments regarding positive aspects of sharing rather than opening data
	Charging for data	Anything to do with payments for data, freemium models
Privacy		
	Breaches	Accidental publishing of data that was in fact personal
	Anonymisation	Awareness of deidentification processes
	Basis for Processing	Reference to 6 bases for processing personal data
Value		
	Internal Value	Of having improved data processes
	Value to Management	In improving and evidencing improvement in city
	Pilot Success	Successful pilots
	Impact	Impact, evaluation of data, etc
	Measuring value	How value of open data is measuring
	Value to Data User	Value of data to piloting companies
	Value to Intermediaries	Value of engaging with OI/data to intermediary organisations

List of References

Aburn, G., Gott, M. and Hoare, K., (2015). What is resilience? An integrative review of the critical literature. *Journal of Advanced Nursing*. **72**(5), 980-1000.

Ahmed, J.U., (2010). Documentary Research Method: New Dimensions. *Indus Journal of Management & Social Sciences.* **4**(1), 1-14.

Almirall, E., Lee, M. and Majchrzak, A., (2014). Open innovation requires integrated competition-community ecosystems: Lessons learned from civic open innovation. *Business Horizons*. **57**(3), 391-400.

Altheide, D., and Schneider, C., (2013). *Qualitative Media Analysis*, Sage Publications, London

Appleton, J.V. and Cowley, S., (1997). Analysing clinical practice guidelines. A method of documentary analysis. *Journal of Advanced Nursing*. **25**, 1008-1017.

Argast, A., and Zvayagintseva, L., (2016). Data Literacy Projects in Canada: Field Notes from the Open Data Institute Node in Toronto. *Journal of Community Informatics Special Issue on Data Literacy*. **12**(3).

Arzberger, P., Schroeder, P., Beaulier, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P. and Wouters, P., (2004). An International Framework to Promote Access to Data. *Science.* **303**(5665), 1777-1778.

Assar, S., Boughzala, I. and Thierry, I., (2011). eGovernment Trends in the Web 2.0 Era and the Open Innovation Perspective: An Exploratory Field Study. *Proceedings of Electronic Government - 10th IFIP WG 8.5 International Conference, EGOV 201.,* Delft, The Netherlands, August 28 - September 2, 2011.

Ayele, W., Juell-Skielse, T., Hjalmarsson, A. and Johanneson, P., (2015). Evaluating Open Data Innovation: A Measurement Model for Digital Innovation Contests. *PACIS 2015 Proceedings*.

Bacon, S., and Goldacre, B., (2020). Barriers to working with National Health Service England's open data. *J Med Internet Res.* **22**(1), e15603.

Bakici, T., Almirall, E., and Wareham, J., (2013). The role of public open innovation intermediaries in local government and the public sector. *Technology Analysis & Strategic Management*. **25**(3), 311-327.

Banterle, F (2018) Data Ownership in the Data Economy: A European Dilemma. *EU Internet Law in the Digital Era.* Springer.

Barry, E. and Bannister, F., (2014). Barriers to open data release: A view from the top. *Information Polity*. **19**(1), 129-152.

Table of Figures

Bauer, M. W., Bicquelet, A. and Suerdem, A. K., (2014). Text Analysis, An Introductory Manifesto In, Bauer, M. W., Bicquelet, A. and Suerdem, A. K., (eds.) (2014) *Textual Analysis. SAGE Benchmarks in Social Research Methods, 1*. Sage, London, UK.

Berends, J, Carrara, W, and Radu, C (2017) Economic Benefits of Open Data. *European Data Portal Analytical Report* 9 [Online]. Luxembourg:European Data Portal. Available from: https://www.europeandataportal.eu/sites/default/files/analytical_report_n9_economic_benefits_of_open_data.pdf

Berg, Bruce L. (2004). *Qualitative Research Methods for the Social Sciences*. (5th edition). Toronto

Boell, S. K. and Cecez-Kecmanovic, D., (2014). A Hermeneutic Approach for Conducting Literature Reviews and Literature Searches. *Communications of the Association for Information Systems*. **34**(12), 257-286.

Boswarva, O., (2015). Global Open Data Index 2015 - United Kingdom Insight. *Open Knowledge International Blog.* [Online]. London:Open Knowledge Foundation. [22/08/16] Available from: http://blog.okfn.org/2015/12/09/Global-Open-Data-Index-2015-United-Kingdom-Insight/

Both, W., (2012). Open Data: What do Citizens Really Want? *Journal of Community Informatics Special Issue on Open Data.* **12**(3).

Boudreau, K., and Lakhani, K., (2009). How to manage outside innovation. *MIT Sloan Management Review*. **50**(4), 69-76.

Bowen, G. A., (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*. **9(**2), 27-40.

Bower, J., L. and Christensen, C., M., (1995). Disruptive technologies: catching the wave. Harvard Business Review, (1995). **73**(1), 43-53.

Boyatzis, R. E., (1998). *Transforming qualitative information: thematic analysis and code development.* Sage Publications, Thousand Oaks, CA.

Brandusescu, A., Iglesias, C. and Robinson, K., (2017). Open Data Barometer 4th Edition. [Online] Washington DC:World Wide Web Foundation. [16/02/20] Available from: https://opendatabarometer.org/doc/4thEdition/ODB-4thEdition-GlobalReport.pdf.

Brandusescu, A. and Lämmerhirt, D., (2019) Issues in Open Data: Measurement. In, Davies, T., Walker, S., Rubenstein, M. and Perrini, F., (eds) (2019). *State of Open Data*. African Minds, Johannesburg, SA.

Braun, V. and Clarke, V., (2013). *Successful qualitative research: A practical guide for beginners.* Sage Publications, London, UK.

Bunting, M. and Lansdell, S., (2019). *Designing decision making processes for data trusts: lessons from three pilots* [online]. London: Open Data Institute. [21/01/20]. Available from: http://theodi.org/wp-content/uploads/2019/04/General-decision-making-report-Apr-19.pdf

Cabinet Office (2011) Making Open Data Real: A Government Summary of Responses. [Online] London: Cabinet Office. [16/02/20] Available from: https://www.gov.uk/government/consultations/making-open-data-real.

Canino, A., (2019). Deconstructing Google Dataset Search. *Public Services Quarterly*. **15(**3), 248-255.

Carnelley, P., Schwenk, H., Cattaneo, G., Micheletti, G., and Osimo, D., (2016). *Europe's Data Marketplaces - Current Status and Future Perspectives*, *European Data Market Study.* [Online]. IDC: Luxembourg. [21/01/20]. Available from: http://datalandscape.eu/data-driven-stories/europe's-data-marketplaces-–-current-status-and-future-perspectives

Carrara, W., Radu, C., and Vollers, H., (2017). Open data maturity in Europe 2017 n3: Open data for a European data economy. [Online]. Brussels: European Data Portal. [21/01/20]. Available from: https://www.europeandataportal.eu/sites/default/files/edp_landscaping_insight_report_n3_2017.pdf.

Chan, C. M. L., (2013). From Open Data to Open Innovation Strategies: Creating E-Services Using Open Government Data. *2013 46th Hawaii International Conference on System Sciences*, Wailea, Maui, HI, 2013, pp. 1890-1899.

Chesbrough, H., (2003). *Open innovation: The New Imperative for Creating and Profiting from Technology.* Harvard Business School Press, Boston, MA.

Chesbrough, H. and Crowther, A.K., (2006). Beyond high tech: early adopters of open innovation in other industries. *R&D Management.* **36**(3), 229-236.

Chesbrough, H.W. and Appleyard, M.M., (2007). Open Innovation and Strategy. *California Management Review*. **50**(1), 57-76.

Chesbrough, H., (2011). Everything You Need to Know About Open Innovation*. Forbes* [online]. (21/03/11). [21/01/20]. Available from: https://www.forbes.com/sites/henrychesbrough/2011/03/21/everything-you-need-to-know-about-open-innovation/

Chesbrough, H. and Bogers, M., (2013). Explicating Open Innovation: Clarifying an Emerging Paradigm for Understanding Innovation. In, Chesbrough, H., Vanhaverbeke, W.and West, J., (eds) (2017). *New Frontiers in Open Innovation*. Oxford University Press, Oxford, UK.

Table of Figures

Chesbrough, H. and Vanhaverbeke, W., (2018). Open Innovation and Public Policy in the EU with Implications for SMEs. In, Vanhaverbeke, W., Frattini, F., Roijakkers, N. and Usman, M., (eds) (2018). *Researching Open Innovation in SMEs*. World Scientific, Hackensack, NJ.

Chiliswa, Z. and Mukutu, L., (2015). *Building Open Data Infrastructure and Strategies for Effective Citizen Engagement.* Open Data Research Symposium, Ottawa, CA 27 May 2015.

Christensen, C., Raynor, M. and McDonald, R., (2015). What is disruptive innovation? *Harvard Business Review.* **93**, 44-53.

Chui, M., Farrell, D. and Jackson, K., (2014). How Governments Can Promote Open Data *Government Designed for New Times* [online]. New York:McKinsey & Co. [16/02/20]. Available from: https://www.mckinsey.com/~/media/mckinsey/industries/public%20sector/our%20insights/how%20government%20can%20promote%20open%20data/how_govt_can_promote_open_data_and_help_unleash_over_$3_trillion_in_economic_value.ashx Issue 2, April 2014 McKinsey & Co, New York

City of London (2016) London City Data Strategy. *Mayor of London & London Assembly*. London: Greater London Authority. [16/02/20] Available from: https://data.london.gov.uk/dataset/data-for-london-a-city-data-strategy

Cohen, B., Almirall, E. and Chesbrough, H., (2016). The City as a Lab: Open Innovation Meets the Collaborative Economy'. *California Management Review*. **59**(1), 5–13.

Conradie, P. and Choenni, S., (2014). On the Barriers for Local Government Releasing Open Data. *Government Information Quarterly*. **31**(1) S10-S17.

Corrales-Garay, D., Mora-Valentin, E-M. and Ortiz-de-Urbina-Criado, M., (2019). Open Data for Open Innovation: An Analysis of Literature Characteristics. *Future Internet.* **11**(3), 77.

Cox, A., Milstead, A. and Gutteridge, C., (2015). *The Organisation Profile Document (OPD) – Simple programming enabling data discovery.* Open Data Research Symposium, Ottawa, CA 27 May 2015

Dahlander, L. and Gann, D.M., (2010). How open is innovation? *Research Policy*. **39**(6), 699-709.

Davies, T., Perini, F. and Alonso, M. J., (2013). Researching the emerging impacts of open data, *ODDC Working Papers #1*, (online). Open Data Research Network. [16/02/20] Available from: http://www.opendataresearch.org/reports

Davies, T., (2010). Open Data, Democracy and Public Sector Reform (online). [16/02/20] Available from: http://www.opendataimpacts.net/report

Davies, T., (2015). Open data in developing countries: Emerging insights from phase 1. *Open Data in Developing Countries Working Papers number 2* (online). Berlin: World Wide Web Foundation.

[21/01/20] Available from: http://www.opendataresearch.org/content/2014/704/open-data-developing-countries-emerging-insights-phase-i.html

Davies, T.G. and Bawa, Z.A., (2012). The promises and perils of open government data (OGD). *The Journal of Community Informatics*. **8**(2).

Debackere, K., Andersen, B., Dvorak, I., Enkel, E,.Krüger, P., Malmqvist, H., Plečkaitis, A., Rehn, A., Secall, S., Stevens, W., Vermeulen, E. and Wellen, D., (2014). Independent Expert Group Report on Open Innovation and Knowledge Transfer Boosting Open Innovation and Knowledge Transfer in the European Union (online). Luxembourg: European Union Directorate-General for Research and Innovation. [16/02/20]. Available from: https://op.europa.eu/en/publication-detail/-/publication/5af0ec3a-f3fb-4ccb-b7ab-70369d0f4d0c

De Cindio, F., (2012). Guidelines for Designing Deliberative Digital Habitats: Learning from e-Participation for Open Data Initiatives. *Journal of Community Informatics Special Issue: Community Informatics and Open Government Data*. **8**(2).

De Freitas, R.K.V. and Dacorso, A.L.R., (2014). Open Innovation in Public Management: Analysis of the Brazilian Action Plan for Open Government Partnership. *Rev. Adm. Pública.* **48**(4), 869–888.

Demeyer, T., Kresin, F., van Oosteren, C. and Gallyan, K., (2012). Apps for Amsterdam. *Journal of Community Informatics Special Edition on Open Government Data.* **8**(2).

Denscombe, M. (1998). *The Good Research Guide for small-scale social research projects.* Open University Press, Buckingham, UK.

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, P., Tait, J. and Zijlstra, T., (2009). *Open data handbook*. [Online]. Cambridge: Open Knowledge International. [16/02/20]. Available from http://opendatahandbook.org.

Dix, A., (2014). Open Data Islands and Communities. *Annals of the Tiree Tech Wave* [Online]. [19/01/17]. Available from: http://tireetechwave.org/wp-content/uploads/2014/07/Open-Data-Islands-and-Communities-v3b.pdf

Domingo, A., Bellalta, B., Palacín, M., Oliver, M. and Almirall, E., (2013). Public open sensor data: Revolutionizing smart cities. *IEEE Technology and Society Magazine*. **32**(4), 50-56.

Denscombe, M. (1998). *The Good Research Guide for small-scale social research projects.* Open University Press, Buckingham, UK

Dodds, L., (2013). Reusers Guide to Open Data Licensing. *Open Data Institute* [Online]. London: Open Data Institute. [21/01/20] Available from: https://theodi.org/article/reusers-guide-to-open-data-licensing/

Table of Figures

Dodds, L, (2016) The State of Open Data Licensing. *Open Data Institute* [Online]. London: Open Data Institute. [21/01/20] Available from*:* https://blog.ldodds.com/2016/06/23/the-state-of-open-licensing/

Dougherty, D. and Hardy, C., (1996). Sustained Product Innovation in Large, Mature Organizations: Overcoming Innovation-to-Organization Problems. *The Academy of Management Journal.* 39**(5)**, 1120-1153.

Duch-Brown, N., Martens, B. and Mueller-Langer, F., (2017). *The economics of ownership, access and trade in digital data; Digital Economy Working Paper* 2016-10 [online]. Seville:JRC Technical Report. [21/01/20]. Available from: https://ec.europa.eu/jrc/sites/jrcsh/files/jrc104756.pdf

Dulong de Rosnay, M. and Janssen, K. (2014) Legal and Institutional Challenges for Opening Data Across Public Sectors: Towards Common Policy Solutions. J. Theor. Appl. Electron.Commer. Res **9(3)**

Dumont, M. and Meeusen, W., (2000). Knowledge spillovers through R&D cooperation. *OECD-NIS Focus Group on Innovative Firms and Networks*, Rome, 15-16 May, 2000.

Eaves, M., (2013). Beyond property rights: Thinking about moral definitions of openness [Online}. *TechPresident* [21/01/20]. Available from: http://techpresident.com/news/wegov/24244/beyond-property-rights-thinking-about-moral-definitions-openness

Eisenhardt, K.M., (1989). Building theories from case study research. *Academy of Management Review.* **14**(4), 532-550.

Eschenfelder, K.R. and Johnson, A., (2014). Managing the Data Commons. *J Assn Inf Sci Tec*. **65**(9), 1757-1774.

European Commission, (2018). *Towards a Common European Data Space* COM 232 [online]. Brussels:EC [21/01/20]. Available from: https://ec.europa.eu/digital-single-market/en/news/communication-towards-common-european-data-space

Fereday, J. and Muir-Cochrane, E., (2006). Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. *Int J Qual Methods*. **5**(1), 80–92.

Ferro, E. and Osella, M. (2013) Eight Business Model Archetypes for PSI Re-Use. *Open Data on the Web Workshop.* 23-24 April 2013, Google Campus, London UK.

Fisher, J. and Fortmann, L., (2010). Governing the data commons: Policy, practice, and the advancement of science. *Information & Management*. **47**(4), 237-245.

Frank, M. and Walker, J., (2016). User Centred Methods for Measuring the Value of Open Data, *Journal of Community Informatics Special Issue on Open Data for Social Change and Sustainable Development.* **12**(2).

Frey, J. and Fontana, A., (1991). The Group Interview in Social Research. *The Social Science Journal*, **28**(2), 175-187.

Frey, K., Luthje, C. and Haag, S., (2011). Whom Should Firms Attract to Open Innovation Platforms? The Role of Knowledge Diversity and Motivation. *Long Range Planning,* **44**(5-6), 397-420.

Fricker, S. and Maksimov, Y., (2017). Pricing of Data Products in Data Marketplaces. In: Ojala A., Holmström, O. and Werder K. (eds) S*oftware Business. ICSOB 2017. Lecture Notes in Business Information Processing, vol 304.* Springer, Champagne, IL.

Fugard, A.J.B. and Potts, H.W.W., (2015). Supporting thinking on sample sizes for thematic analyses: a quantitative tool. *International Journal of Social Research Methodology.* **18**(16), 669–684.

Gaskell, G., (2000). Individual and Group Interviewing. In, Bauer, M. W. and Gaskell, G. (eds) *Qualitative Researching with Text, Image and Sound, A Practical Handbook for Social Research.* Sage Publishing, London, UK.

Gassmann, O., Enkel, E. and Chesbrough, H., (2010). Editorial: The future of open innovation. *R&D Management*. **40**(3), 231–21.

Gerring, J., (2004). What Is a Case Study and What Is It Good for? *American Political Science Review.* **98**(2), 341-354.

Gibbert, M., Ruigrok, W. and Wicki, B., (2008). What passes as a rigorous case study? *Strategic Management Journal*. **29**(13), 1465-1474.

Gomez-Cruz, E. and Thornham, H. (2016). Hackathons, data and discourse: Convolutions of the data(logical). *Big Data and Society*. **3**(2).

Gomer, R., O'Hara, K. and Simperl, E., (2016). Analytical Report 3: Open Data and Privacy. European Data Portal [Online]. Luxembourg: European Data Portal. [16/02/20]. Available from: https://www.europeandataportal.eu/sites/default/files/open_data_and_privacy_v1_final_clean.pdf

Granickas, K., (2014). Building Community Around Open Government Data. ePSI Topic Report.

[Online]. [16/03/16] Available from: http://www.epsiplatform.eu/sites/default/files/Building%20Community%20around%20Open%20Government%20Data.pdf

Gray, J., (2017) *What Do Data Portals Do? Tracing the Politics of Public Information Infrastructures on the Web* Presented at Data Publics, Lancaster University, 31 March – 2 April 2017http://datapublics.net/wp-content/uploads/2017/03/DataPublics-Program-abstracts.pdf

Table of Figures

Grossman, R., Heath, A., Murphy, M., Patterson, M. and Wells, W., (2016). A Case for Data Commons: Towards Data Science as a Service. *Comput Sci Eng.* **18**(5), 10–20.

Gurin, J., (2014) *Open Data Now: The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation.* McGraw Hill, New York, NY.

Gurstein, M., (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday.* **16**(2).

Gutzmer, A., (2016). *Urban innovation networks: Understanding the city as a strategic resource*. Springer, New York, NY.

Haggarty, L., (1996). What is content analysis? *Medical Teacher*. **18**(2), 99-101.

Hall, W. and Pesenti, J., (2017). *Growing the Artificial Intelligence Industry in the UK* [online]. London:Department for Digital, Culture, Media & Sport and Department for Business, Energy & Industrial Strategy. [21/01/20]. Available from: https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk

Hannis, M., (2013). Land Registry: Data for Sale. *The Land [Online].* [22/08/16]. Available from: http://www.thelandmagazine.org.uk/articles/land-registry-data-sal

Hardinges, J. and Wells, P., (2019). Data trusts will not be the final word on data sharing, but they might help. *Public Money & Management.* **39**(5), 320-321.

Harrison, S. (2016). Blockchain and Open Data: How could it work? Open Data Institute Blog [Online]. London: Open Data Institute. [22/08/16]. Available from: http://theodi.org/blog/blockchain-and-open-data-how-could-it-work

Harrison, T.M., Pardo, T.A. and Cook, M., (2012). Creating Open Government Ecosystems: A Research and Development Agenda. *Future Internet*. **4**(4), 900 – 928.

Hellberg, A.S. and Hedström, K., (2015). The story of the sixth myth of open data and open government. *Transforming Government: People, Process and Policy*. ***9***(1), 35-51.

Heimstadt, M. (2014). *The British Open Data Ecosystem.* MSc Dissertation, University of St Andrews.

Hivon, J. and Titah, R., (2015). Citizen Participation in Open Data Use at the Municipal Level*. Open Data Research Symposium.* Ottawa, 27 May, 2015

Hjalmarsson, A. and Rudmark, D., (2012). Designing Digital Innovation Contests, in, Peffers, K., Rothen-berger, M. and Kuechler (eds.): *DESRIST 2012*, LNCS 7286, 9-27.

Hjalmarsson, A., Johannesson, P., Juell-Skielse, G., Rudmark, D., (2014). Beyond Innovation Contests: A Framework of Barriers to Open Innovation of Digital Services. *Proceedings of the 22nd European Conference on Information Systems (ECIS).* Tel Aviv, Israel, 9–11 June 2014.

Hopia, H., Latvala, E. and Liimatainen, L., (2016). Reviewing the methodology of an integrative review. *Scand J Caring Sci*. **30**(4)**,** 662–669.

Hsieh, H.-F. and Shannon, S.E., (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, **15**(9), 1277–1288.

Huber, F., Wainwright, T. and Rentocchini, F., (2018). Open Data for Open Innovation: Managing Absorptive Capacity in SMEs. *R&D Management.* **18**, 1-16.

Hyett, N., Kenny, A. and Dickson-Swift, V., (2014). Methodology or method? A critical review of qualitative case study reports. *International Journal of Qualitative Studies on Health and Well-being.* **9**(1), 23606

International Open Data Charter (IODC), 2015. [16/02/20]. Available from: opendatacharter.net

Irani, L., (2015). Hackathons and the making of entrepreneurial citizenship. *Science, Technology, & Human Values*. **40**(5), 799-824

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

Jeppeson, L. and Lakhani, K., (2010). Marginality and Problem-Solving Effectiveness in Broadcast Search. *Organization Science*. **21(**5), 1016-1033

Jetzek, T., Avital, M, and Bjorn-Andersen, N., (2013). The Generative Mechanisms of Open Government Data. *ECIS 2013 Completed Research.* Paper 156.

Jobin, A., Ienca, M. and Vayena, E., (2019). The global landscape of AI ethics guidelines. *Nat Mach Intell.* **1**, 389–399.

Jones-Devitt, S., Austen, L. and Parkin, H., (2017). Integrative reviewing for exploring complex phenomena. *Social Research Update*, **66**.

Juell-Skielse, G., Hjalmarsson, A., Juell-Skielse, E., Johannesson, P. and Rudmark, D., (2014). Contests as innovation intermediaries in open data markets. *Information Polity*. **19**(3-4), 247-262.

Kankanhalli, A., Zuiderwijk, A. and Kumar Tayi, G., (2017). Open innovation in the public sector: A research agenda. *Government Information Quarterly*. **34**(1), 84-89

Kayser-Bril, N., (2016). Don't ask too much from data literacy. *The Journal of Community Informatics Special Issue on Data Literacy. **12***(3).

Keller, J., (2018), Mapping the Wide World of Data Sharing*, The Open Data Institute*, London [Online]. Available at https://theodi.org/project/the-data-access-map/

Table of Figures

Khoo, C.S.G., Na, J.-C. and Jaidka, K., (2011) Analysis of the Macro-Level Discourse Structure of Literature Reviews. *Online Information Review*. **35**(2). 255-271.

Klievink, B., van der Voort, H. and Veeneman, W., (2018). Creating value through data collaboratives: Balancing innovation and control. *Information Polity.* **23**(4), 379–397.

Koesten, L., Walker, J. and Simperl, E., (2020). Automated Assessment of (Open) Data Portals. European Data Portal [Online}. Luxembourg: European Data Portal.

Kohnstamm, J. and Madhub, D., (2014). Mauritius Declaration on the Internet of Things 36th Annual Conference of Data Protection and Privacy Commissioners, Balaclava, 14 October 2014. [21/01/20] Available from: https://edps.europa.eu/sites/edp/files/publication/14-10-14_mauritius_declaration_en.pdf

Koput, K.W., (1997). A chaotic model of innovative search: some answers, many questions. *Organisation Science.* **8**(5), 528-542.

Koutroumpis, P., Leiponen, A. and Thomas, L., (2017). The (Unfulfilled) Potential of Data Marketplaces, *ETLA Working Papers* 53 [Online]. Helsinki: Etla. [21/01/20]. Available from: https://www.etla.fi/wp-content/uploads/ETLA-Working-Papers-53.pdf

Krishnamurthy, R. and Awazu, Y., (2016). Liberating data for public value: The case of Data.gov. *International Journal of Information Management.* **36**(4), 68–67.

Kröger, J. (2019) Unexpected Inferences from Sensor Data: A Hidden Privacy Threat in the Internet of Things. In, Strous L., Cerf V. (eds) *Internet of Things. Information Processing in an Increasingly Connected World.* IFIPIoT 2018. IFIP Advances in Information and Communication Technology, 548. Springer, Champagne, IL.

Kucera, J. and Chlapek, D., (2013). Comparison of approaches to publication of Open

Government Data in two Czech public sector bodies [Online]. Geneva: World Wide Web Consortium. [16/02/20]. Available from: https://www.w3.org/2013/share-psi/wiki/images/c/c3/Share-PSI_Samos_WS_UEP.pdf

Kucera, J, Chlapek, D, Klimek, J. and Necasky, M. (2015) Methodologies and best practices for open data publication. In *Annual International Workshop on Databases, Texts, Specifications and Objects.* 2015; 1343: 52-64. Available from: http://ceur- ws.org/Vol-1343/.

Lassinantti, J., (2013). Public sector open data: an inside-out open innovation process. In *6th ISPIM Innovation Symposium–Innovation in the Asian Century.* Melbourne, Australia 8-11 December 2013.

Lassinantti, J., (2014). The contradicting view about user groups relation to innovation: the European open data case. In. *Information Systems Research Seminar in Scandinavia: IRIS 37: Designing Human Technologies.* 10-14 August, 2014

Laursen, K. and Salter, A., (2006). Open for innovation: the role of openness in explaining innovation performance among U.K. manufacturing firms. *Strat. Mgmt. J*. **27**, 131-150.

Lee, M., Almirall, E. and Wareham, J., (2016). Open Data & Civic Apps: 1st Generation Failures, 2nd Generation Improvements. *Communications of the ACM.* **59**(1), 82-89.

Lee, S.M., Hwang, T. and Choi, D. (2012). Open innovation in the public sector of leading countries. *Management Decision.* **50**(1), 147-162.

Lee, G. and Kwak, Y.H., (2011). Open Government Implementation Model: A Stage Model for Achieving Increased Public Engagement. In, *Proceedings of the 12th Annual International Conference on Digital Government Research.* College Park, MD, USA,12 - 15 June, 2011.

Leonard, S., (2012). The Fog of More. *The New Inquiry* [Online]. [16/02/20]. Available from: https://thenewinquiry.com/the-fog-of-more/

Leonard-Barton, D., (1987). Implementing Structured Software Methodologies: A Case of Innovation in Process Technology. *INFORMS Journal on Applied Analytics.* **17**(3), 6-17.

Lichtenthaler, U. and Ernst, H., (2009). Opening up the innovation process: the role of technology aggressiveness. *R&D Management*. **39**(1), 38-54.

Lin, A.C., (1998). Bridging Positivist and Interpretivist Approaches to Qualitative Methods. *Policy Studies Journal.* **26,** 162-180.

Lindman, J., Kinnari, T. and Rossi, M., 2014, January. Industrial open data: Case studies of early open data entrepreneurs. In, *System Sciences (HICSS), 2014 47th Hawaii International Conference.* 739-748.

Longshore Smith, M. and Seward, R., (2017). Openness as social praxis. *First Monday*, **22**(4).

Luck, L., Jackson, D. and Usher, K., (2006). Case study: A bridge across the paradigms. *Nursing Inquiry*. **13**(2), 103–109.

Mankins, J. (1995). *Technology readiness levels, A White Paper*, NASA, Washington, DC.

Maguire, M. and Delahunt, B., (2017). Doing a Thematic Analysis: A Practical, Step-by-Step Guide for Learning and Teaching Scholars. *AIJHE.* **9**(3.

Manyika, J., Chui, M.,Groves, S., Farrell, D., Van Kuiken, S. and Almasi Doshi, E., (2013) Open data: Unlocking innovation and performance with liquid information Online San Francisco:McKinsey Global Institute Available from: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information

Table of Figures

Mars, M.M., Bronstein, J.L. and Lusch, R.F., (2012). The value of a metaphor: Organizations and ecosystems. *Organizational Dynamics*. ***41***(4).271-280

Martin, S., Foulonneau M., Turki, S. and Ihadjadene, M., (2013). Risk Analysis to Overcome Barriers to Open Data, *Electronic Journal of e-Government*. **11**(1), 348 -359.

McClean, T., 2011. Not with a Bang but a Whimper: The Politics of Accountability and Open Data in the UK. *APSA 2011 Annual Meeting Paper*.

McDonald, S. and Porcaro, K., (2015). *The civic trust*. [Online}. (04/08/15). [16/02/20]. Available from: https://medium.com/@McDapper/the-civic-trust-e674f9aeab43

McLeod, M. and McNaughton, M., (2016). Mapping an Emerging Open Data Ecosystem. *Journal of Community Informatics Special Issue on Open Data for Social Change and Sustainable Development.* **12**(2).

Mergel, I., Bretschneider, S. I., Louis, C. and Smith, J., (2014). "The Challenges of Challenge.Gov: Adopting Private Sector Business Innovations in the Federal Government." In *Proceedings of the 2014 47th Hawaii International Conference on System Sciences (HICSS)*, 2073–2082. Washington, DC.

Mergel, I., (2018). Open innovation in the public sector: drivers and barriers for the adoption of Challenge.gov. *Public Management Review*. **20**(5), 726-745.

Miles, M.B. and Huberman, A.M. (1994). *Qualitative Data Analysis,* Sage Publications*,* Thousand Oaks, CA

Mladenow, A., Bauer, C. and Strauss, C., (2014). Social Crowd Integration in New Product Development: Crowdsourcing Communities Nourish the Open Innovation Paradigm. *Glob J Flex Syst Manag.* **15**, 77–86.

Moher D., Liberati, A., Tetzlaff, J., Altman, D.G., PRISMA Group, (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med.* **6**(7), e1000097.

Mulgan, G. and Straub, V., (2019). *The new ecosystem of trust: How data trusts, collaboratives and coops can help govern data for the maximum public benefit* [online]. London:Nesta. [21/02/20] Available from: https://www.nesta.org.uk/blog/new-ecosystem-trust/

Mullagh, L. and Walker, J., (2017). Datadrifts: An inclusive route to remote community engagement with open data. *Data Publics*. Lancaster University 31 March – 2 April, 2017

O'Hara, K., (2017). AI in the UK: a short history. In, Hall, W. and Pesenti, J. (eds.) *Growing the artificial intelligence industry in the UK.* London. Department for Digital, Culture, Media and Sport, UK Government. 18-20.

O'Hara, K., (2019). *Data Trusts: Ethics, Architecture and Governance for Trustworthy Data Stewardship* [online]. Southampton: Web Science Institute. [21/01/20]. Available from: https://eprints.soton.ac.uk/428276/1/WSI_White_Paper_1.pdf

Ojo, A., Curry, E. and Zeleti, F., (2015). A Tale of Open Data Innovation in Five Smart Cities. *HICSS '15: Proceedings of the 2015 48th Hawaii International Conference on System Sciences*. January 2015

O'Leary, Z., (2014). *The essential guide to doing your research project* (2nd ed). Sage Publications, Thousand Oaks, CA

Osterwalder, A., (2004). *The business model ontology: A proposition in a design science approach.* PhD. University of Lausanne.

Ovans, A., (2015). What is a business model? *Harvard Business Review* [Online] [16/02/20] Available from: https://hbr.org/2015/01/what-is-a-business-model

Pollock, R., (2013). Open Data Census: Tracking the state of open data around the world. *Open Knowledge Foundation Blog*. (20/02/13) [16/02/20]. Available from: http://blog.okfn.org/2013/02/20/open-data-census-tracking-the-state-of-open-data-around-the-world/

Porway, Jake (2013) You Can't Just Hack Your Way to Social Change, *Harvard Business Review Online.* [16/02/20]. Available from: https://hbr.org/2013/03/you-cant-just-hack-your-way-to/.

Potter, P., (2015). Guaranteeing the Integrity of a Register. *Technology at GDS Blog [*Online]. [22/08/16]. Available from: https://gdstechnology.blog.gov.uk/2015/10/13/guaranteeing-the-integrity-of-a-register/

Remenyi, D., (2011). Field Methods for Academic Research: Interviews, Focus Groups and Questionnaire.s 2nd ed. Academic Conferences Ltd.

Richter, H. and Slowinski, P.R., (2019). The Data Sharing Economy: On the Emergence of New Intermediaries. *IIC.* **50**, 4–29.

Rocher, L. Hendrickx. J. and de Montjoye, Y.-A., (2019) Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*. **10**

Rogawski, C, Verhulst, S and Young, (2016) United Kingdom's Ordnance Survey Open Data: A Clash of Business Models Online at http://odimpact.org/case-united-kingdoms-ordnance-survey-opendata.htmlAccessed 28 July 2016

Rogers, R., (2009). The End of the Virtual - Digital Methods. *Inaugural Speech, Chair, New Media & Digital Culture,* University of Amsterdam, 8 May 2009

Table of Figures

Roman, D. and Gatti, S., (2016). Towards a Reference Architecture for Trusted Data Marketplaces: The Credit Scoring Perspective. *2nd International Conference on Open and Big Data, (OBD) 2016,* Vienna, Austria, August 22-24, 2016

Rossi, F., Russo, M., Sardo, S. and Whitford, J., (2010). Innovation, generative relationships and scaffolding structures: implications of a complexity perspective to innovation for public and private interventions. In, Ahrweiler, P. (ed). Innovation in complex social systems. Routledge Studies in Global Competition. Routledge. Abingdon, UK

Rubenstein, M, Cowls, J and Cath, C (2016) *Opendata.Innovation: An international journey to discover innovative uses of open government data* London: Nesta. [21/03/18]. Available from: https://www.nesta.org.uk/report/opendatainnovation-an-international-journey-to-discover-innovative-uses-of-open-government-data/

Ruijer, E., Grimmelikhuijsen,S. and Meijer A. (2017) Open data for democracy: Developing a theoretical framework for Open Data use. *Government Information Quarterly*, **34** (1) 45-52

Sandelowski, M., (1995). Sample size in qualitative research. *Res. Nurs. Health*, **18**, 179-1.

Sanderson, F., (2013). Investigating Public Participation in Open Government Data. *Open Data Institute Blog*. London: Open Data Institute [16/03/16] Available from: https://theodi.org/blog/open-participation-public-participation-and-open-government-data

Sasse, T., Smith, A., Broad, E., Tennison, J., Wells,P. and Atz, U (2017). *Recommendations for Open Data Portals: From setup to sustainability* [Online]. Luxembourg: European Data Portal [16/02/20]. Available from: https://www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf

Seawright, J. and Gerring, J., (2008). Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options. *Political Research Quarterly*. **61**(2), 294–308.

Scassa, T., (2018). Digital governance and Sidewalk Toronto: Some thoughts on the latest proposal [Online]. [21/01/20] Available from: http://www.teresascassa.ca/index.php?option = com_k2&view = item&id = 290:digital-goverance-and-sidewalk-toronto-some-thoughts-on-the-latest-proposal&Itemid = 80

Schomm, F., Stahl, F. and Vossen, G., (2013). Marketplaces for data: an initial survey. *SIGMOD Rec*. **42**(1), 15–26.

Schöpfel, J., (2011). Towards a Prague definition of grey literature. *The Grey Journal*. **7**, 5–18.

Schwabe, G., (2019). The role of public agencies in blockchain consortia: Learning from the Cardossier. *Information Polity.* **24**(4), 437-451.

Seltzer, E. and Mahmoudi, D., (2012). Citizen Participation, Open Innovation, and Crowdsourcing: Challenges and Opportunities for Planning. *Journal of Planning Literature.* **28**(1), 3-18.

Shadbolt, N., (2010). Towards a pan-EU Data Portal - data.gov.eu [Online] (31/08/10). [30/08/17] Available from: http://ec.europa.eu/digital-agenda/en/open-data-portals

Shekhar, S. and Canares, M., (2016). Open Data and Sub-national Governments: Lessons from Developing Countries. *Journal of Community Informatics, Second Special Issue Community Informatics and Open Government Data.* **12**(2).

Sinkovics, N., (2018). Pattern matching in qualitative analysis. In, Cassell, C., Cunliffe, A.L. and Grandy, G. (eds) (2018). *The SAGE handbook of qualitative business and management research methods.* Sage Publications, Thousand Oaks, CA

Śledzik K., (2013), Schumpeter's view on innovation and entrepreneurship, *Management Trends in Theory and Practice*, (ed.) Stefan Hittmar, Faculty of Management Science and Informatics, University of Zilina & Institute of Management

Smith, G. and Sandberg, J., (2018). Barriers to Innovating with Open Government Data: Exploring Experiences across Service Phases and User Types. *Inf. Polity*. **23**(3), 249–265.

Snyder, H., (2019). Literature Review as a Research Methodology: Overview and Guidelines. *Journal of Business Research*. **104**, 333-339.

Sørensen, E. and Torfing, J. (2011). Enhancing collaborative innovation in the public sector. *Administration & Society.* **43**(8), 842–868.

Sollazzo, G, and Miller, D. (2017) *Open Data in the Health Sector* [Online]. [25/02/17]. Available from: http://openhealthcare.org.uk/open-data-in-the-health-sector

de Souza, M., da Silva, M. and de Carvalho, R., (2010). Integrative review: what is it? How to do it?, *Einstein (São Paulo)*. **8**(1), 102-106.

Stahl, F., Schomm, F., Vossen, G. and Vomfell, L., (2016). A Classification Framework for Data Marketplaces. *Vietnam J Comput Sci.* **3,** 137-143.

Stahl, F., Schomm, F., Vomfell, L. and Vossen, G., (2017). Marketplaces for Digital Data: Quo Vadis? *Computer and Information Science.* **10**(4), 22-37.

Stalla-Bourdillon, S., Wintour, A. and Carmichael, L., (2019). *Building Trust Through Data Foundations* [online]. Southampton: Web Science Institute. [21/01/20]. Available from: https://cdn.southampton.ac.uk/assets/imported/transforms/content-block/UsefulDownloads_Download/69C60B6AAC8C4404BB179EAFB71942C0/White%20Paper%202.pdf

Stalla-Bourdillon, S., Thuermer, G., Walker, J., and Carmichael, L., (2020). Data Protection by Design: Building the foundations of trustworthy data sharing. *Data & Policy,* **1.**

Stallman, R., (1985). The GNU Manifesto. *Dr. Dobb's Journal.* **10**(3), 30.

Susha, I., Johannesson, P., and Juell-Skielse, G., (2016). Open Data Research in the Nordic Region: Towards a Scandinavian Approach? In, *Proceedings of the 15th IFIP WG 8.5 International Conference, EGOV 2016.* Guimares, Portugal, September 2016

Susha, I., Janssen, M. and Verhulst, S., (2017). Data Collaboratives as a New Frontier of Cross Sector Partnerships in the Age of Open Data: Taxonomy Development. In, *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2691–2700.

Susha, I., Janssen, M. and Verhulst, S., (2017a). Data collaboratives as "bazaars"?: A review of coordination problems and mechanisms to match demand for data with supply. *Transforming Government: people, process and policy*. (**11**)1, 157-172.

Susha, I., Pardo, T., Janssen, M., Adler, N. and Verhulst, S., (2018). A Research Roadmap to Advance Data Collaboratives Practice as a Novel Research Direction. *International Journal of Electronic Government Research*. **14**(3), 1-11.

Teece, D. J., (1986). Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy.* **15**(6), 285-305.

Temiz, S., Bogers, M. and Brown, T.E., (2019). The Ecosystem of Open Data Stakeholders in Sweden. *Digital Innovation.* 79-112.

Tennison, J., (2012). *Open Data Business Models*

[Online]. [19/03/16]. Available from: http://www.jenitennison.com/2012/08/20/open-data-business-models.html

Tidd, J., (2013). Why We Need a Tighter Theory and More Critical Research on Open Innovation

In Tidd, J., (ed) (2013) *Open Innovation Research, Management and Practice*. World Scientific, Hackensack, NJ.

Tidd, J. and Bessant, J., (2013). *Managing Innovation: Integrating Technological, Market and Organizational Change*. 5th ed. John Wiley & Sons, Chichester, UK.

Tong, A., Irshad, H. and Revell Ward, D., (2013). *Market Assessment of Public Sector Information* [Online]. London: Department of Business, Skills and Innovation. [16/02/20]. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/198905/bis-13-743-market-assessment-of-public-sector-information.pdf

Torraco, R.J., (2005). Writing Integrative Literature Reviews: Guidelines and Examples. *Human Resource Development Review.* **4**(3), 356-367.

Torraco, R.J., (2016). Writing Integrative Literature Reviews: Using the Past and Present to Explore the Future. *Human Resource Development Review*. *15*(4), 404–428.

Trott, P. and Hartmann, D., (2009). Why 'open innovation' is old wine in new bottles. *International Journal of Innovation Management.* **13**(4), 715–736.

Ubaldi, B., (2013). Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Working Papers on Public Governance, No. 22*. OECD Publishing.

Vagharseyyedin, S.A., (2016). An integrative review of literature on determinants of nurses' organizational commitment. *Iran J Nurs Midwifery Res*. **21**(2) 107–117.

Van de Vrande, V., de Jong, J.P.J. and Vanhaverbeke, W., (2008). Open Innovation SMEs: Trends, Motives and Management Challenges. *Technovation.* **29**(6-7), 423-437.

van Loenen, B., (2018). Towards a User-Oriented Open Data Strategy. In: van Loenen B., Vancauwenberghe G., Crompvoets J. (eds) *Open Data Exposed. Information Technology and Law Series*, 30 T.M.C. Asser Press, The Hague

Van Schalkwyk, F., Canares, M., Chattapadhyay, S. and Andrason, A., (2015). *Open Data Intermediaries in Developing Countries. Open Data Research Symposium*. Ottawa, CA 27 May 2015

Van Schalkwyk, F., Willmers, M. and McNaughton, M., (2015). Viscous Open Data: The Roles of Intermediaries in an Open Data Ecosystem. *Information Technology for Development* 22:sup1, 68-83.

Van Zoonen, L., (2016). Privacy concerns in smart cities. *Government Information Quarterly*.

33(3), 472-480

Verhulst, S. and Sangokoya, D., (2014). Mapping the Next Frontier of Open Data: Corporate Data Sharing. *Internet Monitor 2014: Data and Privacy*.

Vetrò, A., Canova, L., Torchiano, M., Orozco Minotas, C., Iemma, R. and Morando, F., (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*. **33**(2), 325-337.

Van Loenen, B., Janssen, K. and Welle Donker, F. (2012) Towards true interoperable geographic data: developing a global standard for geo-data licenses. In, Janssen, K. and Crompvoets, J., *Geographic data and the law. Defining new challenges*. (2012) Leuven: Leuven University Press

Van Loenen, B., (2018). Towards a User-Oriented Open Data Strategy. In, Van Loenen, B., Vancauwenberghe, G., and Crompvoets, J., (eds) *Open Data Exposed. Information Technology and Law Series*, vol 30. T.M.C. Asser Press, The Hague, NL

Van Zoonen, L., (2016). Privacy concerns in smart cities. Government Information Quarterly. **33**(3), 472-480.

Von Lucke, J., (2014). Open Societal Innovation, *OpenSym '14,* August 27 - 29 2014, Berlin.

Table of Figures

Walker, J., (2014). *An Investigation of the Perception of the Challenges of Using Open Data in the UK Amongst Stakeholders in Early Stage Companies*. MSc Dissertation, University of Southampton

Walker, J., Taylor, J. and Carr, L. (2015) From Public Sector Information Catalogue to Productive Data: Defining a National Information Infrastructure. *Proceedings of the 3rd International Workshop on Building Web Observatories*, Oxford, UK, 28 June 2015

Walker, J. and Simperl, E (2017) The Future of Open Data Portals. *European Data Portal Analytical Report 8 [Online]. Luxembourg: European Data Portal. [20/01/20] Available from:* https://www.europeandataportal.eu/sites/default/files/edp_analyticalreport_n8.pdf

Walker J. and Simperl, E., (2018) Open Data and Entrepreneurship. *European Data Portal Analytical Report 10* [Online]. Luxembourg: European Data Portal. [20/01/20] Available from: https://www.europeandataportal.eu/sites/default/files/analytical_report_10_open_data_and_entrepreneurship.pdf

Walker, J., Hewitt, S., and Simperl, E., (2020). Assessment of Funding Options for Open Data Portals. *European Data Portal*. Luxembourg: European Data Portal.

Walker, J, Thuermer, G and Simperl, E., (2019). Enabling Smart Rural: The Open Data Gap *European Data Portal Analytical Report 14* [Online]. Luxembourg: European Data Portal. [20/01/20.] Available from: https://www.europeandataportal.eu/sites/default/files/analytical_report_14_enabling_smart_rural.pdf

Walker, J., Ibanez, L. D. and Simperl, E., (2019) Citizen-centric factors of smart cities [Online]. [21/01/20]. Available from: http://smartcityinnovation.eu/wp-content/uploads/2019/06/7389-Final-Smarter-Cities-web-B.pdf

Walker, J., Simperl, E. and Carr, L., (2019). A Framework for Data Sharing for Open Innovation. *17th International Open and User Innovation Conference,* Utrecht, 8-10 July 2019.

Walker, J, Simperl, E, Ibanez, LD, Frank, M, Costa, E, Koesten, L, West P and Hewitt, S (2020) Sustainability of Open Data Portals. *European Data Portal.* Luxembourg: European Data Portal.

Webster, J. and Watson, R.T., (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*. **26**(2), 12 -23.

Welle Donker F, van Loenen, B., (2016). Sustainable Business Models for Public Sector Open Data Providers. *JeDEM eJournal of eDemocracy and Open Government***. 8**(1). 28–61.

Welle Donker, F. and van Loenen, B., (2017). How to assess the success of the open data ecosystem? *International Journal of Digital Earth*. **10**(3), 284-306.

West, J. and Bogers, M., (2017). Open innovation: current status and research opportunities. *Innovatio*n. **19**(1), 43-50.

Whittemore, R. and Kraft, K., (2005). The Integrative Review: updated methodology. *Journal of Advanced Nursing.* **52**(5), 546 - 553.

Worthy, B., (2013). Where Are The Armchair Auditors? *The Open Data Institute Blog [Online].* London: Open Data Institute. [21/01/20]. Available from: https://theodi.org/blog/guest-blog-where-are-armchair-auditors

Yin, R. K., (1994). *Case study research: design and methods.* (2nd edition) Sage, Thousand Oaks.

Yin, R. K., (2003). *Case study research, design and methods* (3rd edition). Sage, Thousand Oaks.

Yin, R. K., (2011). *Qualitative research from start to finish.* The Guilford Press, New York, NY

Yiu, C., (2012). The big data opportunity. *Policy exchange*, **8**

Young, M., Rodriguez, L., Keller, E., Sun, F., Sa, B., Whittington, J. and Howe, B., (2018). Beyond Open vs. Closed: Balancing Individual Privacy and Public Accountability in Data Sharing. In: *Proceedings of ACM (FAT'19)*. New York, NY: ACM.

Zeleti, F. A., Ojo, A., and Curry, E., (2014). Emerging Business Models for the Open Data Industry: Characterization and Analysis. In, *Proceedings of the 15th Annual International Conference on Digital Government Research,* Mexico, June 2014

Zimmerman, H. and Pucihar, A., (2015) Open Innovation, Open Data and New Business Models. In Doucek, P., Chroust, G. and Oskrdal, V.(eds.), *Proceedings of IDIMT 2015 - 23rd Interdisciplinary Information and Management Talk*s, Poděbrady, Czech Republic,September 9-11, 2015.

Zuiderwijk, A., Jannsen, M., Choenni, S., Meijer, M. and Alibaks, R. (2012). Socio-Technical Impediments of Open Data. *Electronic Journal of e-Government.* **10**(2), 156-172.

Zuiderwijk, A. and Janssen, M., (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*. ***31***(1), 17-29.