AALTO UNIVERSITY
School of Science
Department of Industrial Engineering and Management

Jingwei Liu

BIG DATA STRATEGIES IN AN OPEN INNOVATION CONTEXT

Bachelor's thesis

Espoo, 06.05.2017

Supervisor: Eerikki Mäki

Thesis advisor: Mona Roman

Aalto University
School of Science

Aalto University School of Science Department of Industrial Engineering and Management		BACHELOR'S THESIS ABSTRACT
Author: Jingwei Liu		
Title: Big Data Strategies in an Open Innovation Context		
Pages: 21 + 5	Date: 06.05.2017	Publishing language: English
Major: Industrial Engineering and Management		Major code: SCI3025.kand

Big data can provide immense economic, scientific and social value. New value and information can be derived from big data by linking up existing data sets. This has pushed organisations to pursue open data initiatives. These initiatives can be found among government, industry and academia.

While organisational benefits are clear to pursue open innovation in the field of big data, individual researchers' motivation for opening their big data sets have not been addressed to make open science realise its true potential.

Scientists' data sharing behaviour is largely driven by perceived career benefits and risks, effort needed to share data and the availability of data repositories. These are factors that need to considered when pushing open data initiatives in academia.

Opportunities for open big data through open innovation collaboration include organisational transparency, accelerated and reproducible research and new businesses. These opportunities are hindered by key stakeholder's unwillingness to participate in collaboration due to perceived risks, privacy and ethical issues and technical issues related to the complexity of big data.

Commercialisation of big research data can be realised through patenting, licencing and spin-outs. There exist data-driven business models that can potentially applied to the research big data of scientists. Most of these business models rely heavily on external data sources. Virtual research environments and boundary organisations are two examples of potential big data ecosystems fostering open innovation collaboration.

Instructor: Mona Roman
Supervisor: Eerikki Mäki
Keywords: big data, open innovation, open science, open data, open research data, triple helix model, technology transfer, motivation, data-driven business model

Table of Contents

Preface

As part of my work at Software Business Lab (SBL) at the Department of Industrial Engineering and Management in Aalto University, this Bachelor's thesis will be a key deliverable in Science2Society (S2S), a project funded by the European Commission under its funding programme Horizon 2020. The aim of S2S is to improve Europe's innovation throughput by examining current ways for technology transfer and providing best practices and guidelines for stakeholders of S2S. Key schemes currently used to encourage use of innovations are studied in the form of seven separate pilots. This Bachelor's thesis will be literature study report part of Pilot number five "Collaboration through Big data and Science 2.0". I would like to thank my employer, Timo Nyberg and my thesis advisor, Mona Roman for providing me with the opportunity and guidance to work on this subject.

Espoo, May 2017

Jingwei Liu

1 Introduction

In recent years, the hype around big data has seen it become a buzzword for many organisations. A reason for this hype can be attributed to the perception that big data is valuable. For example, McKinsey & Company (2011) estimate that big data can have a potential value of up to $300 billion annually the US health care sector and €250 billion annually to Europe's public sector – a number that is higher than the GDP of Greece at the time of the report.

The sheer size of available data has grown immensely over the last few years and today the volume of data is measured in zettabytes – a measure equal to one trillion gigabytes. It is said that about 90% of available digital data is created in the last 2 years. The hype surrounding big data comes from being able to analyse vastly complex big data sets and derive value out of them (Alharthi, Krotov and Bowman, 2017). With big data and big data analytics, platforms like Amazon can recommend products based on customers' buying patterns (Kim, 2014). Similarly, social media websites such as Facebook can use predictive analysis for marketing purposes (Hung, 2016).

Big data itself as data still needs to be processed and transformed into information. This information when analysed can turn into knowledge which can be of value. Different big data sources can also be combined to derive patterns and new information. Researchers from universities and research organisations have their own existing research big data that would enable people to develop new forms of policy, businesses and applications which are data driven. However, useful big data is limitedly available at large scale due to scientists perceiving little benefit of opening their research data – which is their source of scientific reputation.

1.1 Research objective and questions

The objective of this study is to examine whether it is possible to define the relationship between big data and open innovation by deriving a hypothesis between these two from literature. By searching for a link between big data and open innovation, two highly significant phenomena in modern-day academia and industry, we can provide an answer to the following research questions:

1. How to motivate big research data providers to open their data?

2. What are the opportunities and obstacles for current big data strategies to boost open innovation?

3. What are the existing big data strategies?

The purpose of the first research question is to form an understanding on what makes researchers willing or unwilling to open their data sets. The second research question assesses the opportunities and obstacles in opening big data sets as an open innovation practice. The final research question explores what models and best practices are currently used to capture value from big data repositories.

1.2 Scope and research method

This study is conducted as a literature review. The scope is narrowed down to technology transfer from universities and research institutes to industry and the commercialisation of research data. The study will inspect the possibilities to open the data from researchers at these institutes by looking in literature for current strategies deployed by big data repositories and the related best practices and operational models. Literature is chosen based on key word searches ("Open Innovation", "Big Data", "Technology Transfer", "Open Science" etc.) in relevant literature databases (Scopus, Web of Science and Google Scholar). The articles chosen are published after Chesbrough's article *"The Era of Open Innovation"* (2003) due to that work's significance for collaborative work as well as big data being a relatively new phenomenon.

1.3 Key definitions

The terms *big data*, *open innovation* and *open science* are key terms in the context of this study and will be described in this chapter.

1.3.1 Big data

In a sense, the term big data itself is troublesome as the most relevant characteristic of big data is not the sheer volume of a dataset but rather its relationality to other data. Big data is also seen as a analytic phenomenon with the relationality being presented through its ability to exert patterns where none exist (Boyd and Crawford, 2011).

It is difficult to give a clear, uniform definition of big data due a range of existing definitions often having common and altering elements (Emmanuel and Stanier, 2016). The "3Vs" (Volume, velocity, variety) approach, proposed by Laney (2001) to define

challenges in data management of growing modern data, is often used to define the characteristic of big data. Many following authors have expanded on the "3Vs" approach by adding additional features. Following their study on definitions found in existing literature De Mauro, Greco and Grimaldi (2016) define big data as the following:

> *Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value.*

In other words, big data can be understood as a set of largely unstructured data generated at a fast pace and whose size is beyond the analysis capability of conventional database management systems. In addition to this, there also exists an understanding that big data and its applications have the potential to impact society or organisations as well as create value out of information obtained from it.

1.3.2 Open innovation

Open innovation is a term first coined by Chesbrough (2003) in an article on the shifting paradigm in companies' R&D models from a closed to open toward the end of the 20th century. The article focuses on the commercialisation of externally and internally generated ideas as well as pathways to markets for value creation. On one hand, this shift in approach is attributed to the increasing staff mobility. This makes retaining knowledge in-house more difficult but also enables knowledge to flow between organisations. On the other hand, researchers have more options available to pursue new discoveries outside of traditional corporate research labs, e.g. seeking for private capital to a startup or licencing agreements. Chesbrough gave a label for this paradigm, however the concept that organisations make use of external links for knowledge and information to progress technologically dates back to findings from the 1960s (Trott and Hartmann, 2009).

The concept of open innovation can also be applied to academia and builds on the same idea of making the boundaries between environment and organisation porous. In many fields of science research is conducted through collaborative efforts with results indicating that high scientific impact correlates with wide collaboration. (Tacke, 2010)

1.3.3 Open Science

Many scientific findings can be considered the front-end of the innovation process and openness in academia and industrial science can be described with the term *open science*.

Open science fosters collaboration and improves the overall scientific system by sharing discoveries in the earliest stage among interested practitioners. However, unlike open innovation in the business centric view, clear incentives to pursue open science on an individual level are somewhat lacking. (Friesike *et al.*, 2014)

Looking at the entire innovation process universities and research institutes form a crucial pair in the *knowledge transfer exchange* processes through knowledge sharing and cooperation. Directly involved in this process are the individual scientists in these organisations carrying the knowledge. This makes assessing the motivation and factors influencing engagement of researchers highly relevant in knowledge transfer exchange as well as the open innovation prospect between academia and industry. (Padilla-Meléndez and Garrido-Moreno, 2012)

1.4 Outline of the thesis

The thesis is structured as follows. After this introduction, the second chapter presents the current landscape on open big data through the dimensions of industry, government, academia. Following this, the third chapter assesses the incentives and expected return of big research data providers for opening their data. The fourth chapter examines the underlying opportunities and obstacles that hinder researchers and organisations to open all their big data. The chapter that follows explores different forms of models for big data commercialisation and collaboration. Finally, the last chapter summarises the work and investigates the possibility to develop a hypothesis between big data and open innovation.

2 Current landscape

This chapter presents a brief overview of the key stakeholders related to open innovation and their current open big data principles as well as models used for its commercialising. The key stakeholders are government, industry and academia.

The triple helix model suggested by Etzkowitz and Leydesdorff (1995) is often used to conceptualise the university-industry-government relationship to generate research and knowledge-based innovations. In the triple helix model the three involved entities form a triad for the development of innovations and economic value in today's Knowledge Society. In the context of open innovation, the triple helix model also has application in the field of big data. For each dimensions there exists something to be gained from this relationship: novelty production in science, wealth generation in economy and normative control in governance (Park, 2014).

2.1 Government

Open government data refers to non-personal data generated and produced by governments that can be freely used, reused and distributed by anyone. Open government data should be machine-readable and is provided in a raw format (Zimmermann and Pucihar, 2015). Governments release their data to support transparency and encourage the entrepreneurial use of the data for societal impact (Okamoto, 2017).

Zimmermann and Pucihar (2015) state that high capacity countries have adopted open data policies with strong political backing with an aim to have open data on all different government levels. However, capacity constrained countries face challenges in establishing sustainable open data initiatives due to limits of government, civil society or private sector capacity. For many governing bodies making data open and accessible has been on the agenda (Zimmermann and Pucihar, 2015). Data sets already available for download include government data from US at https://www.data.gov/ and from UK at https://data.gov.uk/ (Hand, 2013).

Another example of open government data initiative is the European Commission (EC). The EC has launched the *Open Data Pilot* for selected Horizon 2020 projects as well as all projects starting from 2017 to make their generated data accessible (OpenAIRE, 2016). Horizon 2020 is the European Union's largest funding programme to date with an estimated funding of €80 billion. For project participants, the EC recommends the construction of a living document updated throughout the project in the form of a *Data*

Management Plan (DMP) for their publicly funded projects. The DMP can help to identify challenges in data sharing and future reuse and provides a way overcome them.

Even though open government activities are not inherently profit oriented, the provision of open government data can lead to new and innovative business models. As the data provided is in a raw format, it needs to be processed by government or third parties for value creation. The businesses that develop new business models based on this open data will have a profit oriented focus (Zimmermann and Pucihar 2015).

2.2 Industry

Some industries like insurance and retail companies have traditionally relied on their data to form business decisions. With the emergence of big data, companies have also begun to make use of external data sources such as social media or sensors to capture value by combining data sets. This has led to the development of new data-related ventures and data-driven business models (Hartmann *et al.*, 2016).

Commercialisation of big data can provide a multitude of benefits for organisations ranging from new business insights, improved operating processes as well as faster and better decision making. Big data commercialisation can be made possible through utilisation of external big data, that is often available in the form of open data. An example for exploiting open external data is for companies to the use open government data in conjunction with their own data (Thomas and Leiponen, 2016). From within, some organisations also establish data science methods by, for example, appointing a Chief Data Officer and setting up data science teams to support open innovation approaches. Companies' short-term goal for big data can be understanding customer behaviour from gathered data while in the long-term the goal is to have prediction capabilities (Drexler *et al.*, 2014).

In their article addressing the commercialisation of public research under open innovation models, Cervantes and Meissner (2014) unveil the interests for industry side to collaborate with other organisations. Due to high competition companies are often motivated to form partnerships with other companies, universities or research organisation to complement their innovation activities, essentially applying an open innovation approach and using the triple helix model in their path to market. Their study indicates that this has an impact on both the company's innovation output as well as the scientist's scientific contribution. In their research, Cervantes and Meissner (2014)

suggest that engaging in technology transfer has a positive effect on both the companies' innovation contribution as well as the scientific work, with patenting activity positively affecting the publication output and citation record of researchers.

2.3 Academia

The scientific community has been adopting open science and open data principles for scientific practices. This means allowing the practicing of science so that others may collaborate and contribute by making all research data and processes freely available. This open science phenomenon and open data ecosystems are still at an early stage (Sadiku, Tembely and Musa, 2016). Perception on data sharing differs among disciplines and while many funding agencies require a DMP, as in the case of EC's Horizon 2020 projects, to showcase the possibility of data reuse, researchers tend be cautious with sharing their data (Pampel and Dallmeier-Tiessen, 2014). Evidence that suggest openly sharing data is beneficial for researchers and tools that support open data sharing processes are growing in number. However, a widespread adoption of these open practices have yet to be achieved (McKiernan *et al.*, 2016).

An example to boost open data and exploitation of big data benefits is the *European Cloud Initiative* (EDI) launched by the EC. The aim of this initiative is to boost Europe's data-driven innovations and competitiveness by providing a world-class virtual environment for researchers and science and technology professionals to store and manage their data. The EDI focuses initially on the scientific community with the intended user base expanding to the public sector and industry later. The EDI would make it possible to move, share and reuse big data seamlessly across markets and borders to foster open innovation (European Commission, 2016).

While organisations push out initiatives supporting open data practices, Viseur (2015) lists several reasons that draws the interest of scientists in engaging with open data as well as opening their own data. By sharing experimental data, research results can be reproduced and potentially lead to new developments in that field of research. The emergence of online communities and the semantic Web long linked data are also noted as points of interest for scientists. By sharing data online of unfinished works scientists can accelerate the process of discovery and get feedback on the conducted research. Viseur's (2015) perspective on open research data falls within the context of science 2.0 and open science. In other words, it refers to using Web 2.0 tools and practices while adopting the mind-set of openness and sharing.

2.4 Evaluation

Big data and its potential economic impacts are well documented (McKinsey & Company, 2011). This has pushed organisations to pursue open data initiatives. Ciancarini, Poggi and Russo (2016) note that open data has been a joint interest for public institutions and private companies since 2009.

The possibilities of digitalisations have allowed for big data to be stored, distributed and subsequently analysed by organisations and individuals. Collaborations between different entities are also made possible through a plethora of open data initiatives that are being pushed around the world by governments, institutions and even businesses. However, there exists a general feeling that these open data initiatives have not yet realised their expected potential (Jetzek, Avital and Bjorn-Andersen, 2014).

3 Scientists' motivation for opening data

This chapter looks at what motivates individual scientists to open their data. In other words, it explores scientists' incentives for opening data as well as factors preventing them from opening their data. The chapter examines the both positive and negative drivers that affect scientists' data sharing behaviour.

There are many perceived benefits to big data sharing for companies, governments and academia from a value creation aspect. For universities, research organisations along with the scientific community, the increased transparency of quantitative analytic work, credibility and reproducibility of research are highly valued (Kim and Adler, 2015). While these benefits are what drive organisations to pursue data sharing, it is also important to examine what incentives individual scientists, who produce the data or are owners of data repositories, perceive in sharing it.

3.1 Positive drivers

Personal motivations, e.g. perceived career benefits and risks, expected effort and personal attitude towards data sharing, are primary drivers for scientists' data sharing behaviour (Kim and Adler, 2015; Kim and Zhang, 2015).

In their research, Kim and Zhang (2015) notice that a STEM researcher's attitude towards data sharing directly influences their data sharing behaviour. In other words, positive attitude on data sharing would lead to more data sharing. They identify the factors that positively influencing researcher's attitude towards data sharing as perceived career benefit and normative influence. Career benefits are considered academic rewards like recognition and reputation while normative influence is a researcher's perception on whether others in their field think he or she should share their research data.

Another factor significantly affecting researcher's attitude toward data sharing and data sharing behaviour is the perceived availability of data repositories. This is important since good, uniform community norms for data sharing are yet to be deployed by the scientific community. To develop researcher's data sharing behaviour, the scientific community needs to develop standardised data repositories, make it available and promote them to researchers. (Kim and Zhang, 2015)

3.2 Negative drivers

Factors that negatively influence scientists' attitude towards data sharing are perceived career risks and effort required to share data. This means a researcher believes data sharing can possibly have undesirable consequences on their career or data sharing requires valuable work and time from their part. This is time they spent on making data available in a viable format that could be spend on something that has direct scientific contribution (Kim and Zhang, 2015).

In their survey on 361 social scientists, Kim and Adler (2015) find that funding agencies' and journals' pressure as well as the availability of data repositories are not considered significant factors influencing scientists' data sharing behaviour. Meanwhile, a similar research by Kim and Zhang (2015) conducted with the help of 1298 responses from STEM (science, technology, engineering and mathematics) researchers, show that the aforementioned factors have a significant influence on researchers from these disciplines. This finding is echoed in a paper on data sharing by Tenopir et al. (2011). This is due to only some journals having specific guidelines requiring data sharing with other researchers.

In their paper, Tenopir *et al.* (2011) make note that researchers are reluctant to share their data due to concerns with legal issues, misuse of data, and incompatible data types. This issue is linked to researchers' perception on available data repositories and guidelines for data sharing not being compatible with the complexity of their data. This can lead to misinterpretation in their data or may be used in other ways than intended.

3.3 Key findings on factors influencing data sharing

There are perceived benefits for researchers participating in open science practices (McKiernan *et al.*, 2016). As evidenced by Viseur (2015), this includes open publications gaining more citations and attention as well as reproducibility of open research boosting a scientist's credibility. In addition, there are also perceived risks from scientists' perspective for data sharing. These include perceived effort and negative career influences from data sharing. Benefits and risks are factors that affect scientists' attitude on data sharing which in turn affects their data sharing behaviour (Kim and Adler, 2015; Kim and Zhang, 2015).

A perceived lack of available data repositories and clear guidelines for data storing can also negatively affect their data sharing behaviour. To encourage data sharing behaviour

in scientists, organisations and the scientific community need to understand the personal motivation for scientists to participate in open science and find ways to address them. The positive and negative factors influencing scientists' motivations are presented in table 1 below.

Table 1: Factors affecting data sharing behaviour (Kim and Adler, 2015; Kim and Zhang, 2015)

Positive factor	**Negative factor**	**Varies among fields**
Boosted reputation through data sharing	Perceived career risks	Normative influence (others perception on whether data should be shared)
Academic rewards through data sharing	Perceived effort in making data sharable	
Availability of repositories for data sharing	Unavailability of repositories	

Thomas and Leiponen (2016) note that organisations progress through stages in their big data commercialisation activities. Organisations opening their data for commercialisation typically first experiment with certain quantity of data. With perceived success, they move up the value chain and start collaborating with more partners and suppliers. Once comfortable with the notion of commercialising data, they will begin releasing more data and move from simple data supply to more complex business models. These complex business models generate more revenue but are harder to execute, prompting more open collaboration and co-creation.

This progression of organisations could also apply to individual researchers with data repositories that have interest in commercialising it. However, this would require them also to be willing to open their data first place where their attitude towards data sharing would have to be positive.

4 Opportunities and obstacles in opening big data

In this chapter the underlying opportunities and obstacles for opening big data to boost open innovations approaches are assessed.

4.1 Opportunities

In the era of internet, researchers have less restrictions than before to share their data, especially in a raw format (Kim and Zhang, 2015). According to Liao (2015) the rise in prominence of big data means new opportunities will arise for universities. Liao (2015) argues it is important to integrate business and technology when considering the big data opportunities in higher learning institutes.

By having open access to data, the rate of scientific discovery is accelerated in different research fields (Sadiku, Tembely and Musa, 2016). Aside from increasing government and research transparency open data can also have potential economic value in improved public service at a lower cost and value for society and businesses through accessing and combining data in new ways (Cowan, Alencar and McGarry, 2014). In addition to economic benefits, open research data can also support public policy-making when integrated with open government data (Zuiderwijk *et al.*, 2016). Researchers can also use open research practices to gain more citations, media attention, potential collaborators as well as job and funding opportunities (McKiernan *et al.*, 2016).

In his paper, Hand (2013) argues that the economic growth driven by open data initiatives are more subtle. The author lists accountability and empowering communities as two key enablers of open big data. Both factors are related to transparency of conducted work and people being able to see where actions are needed and how effective they are. The author indicates that people are inherently not interested in the data but rather want answers from the data available. Value from the data comes from the fact that it can be processed and lead us to these answers. Opening access to data provides anyone with the opportunity to generate value out of the data (Jetzek, Avital and Bjorn-Andersen, 2014).

There are some advantages for industry to look for university intellectual property (IP) for innovation purposes. The advantages suggested by Minshall, Seldon and Probert (2007) in licensing university IP to lay in the low investment, potential for multiple revenue streams and limited need to use complementary resources. While licensing university is one option, the creation of a spin-out firm brings the opportunity to capture

a high proportion of generated value and building the entrepreneurial image of the university.

Presented above are reasons for opening big data and establishing big data ventures with external stakeholders. Next the obstacles for opening big data and big data collaborations are introduced.

4.2 Obstacles

Minshall, Seldon and Probert (2007) identify the challenge for industry adopting the open innovation model as the significant investment and time required to generate value out of the university IP due to its low technology readiness. According to Minshall, Seldon and Probert (2007) the cons for licencing university IP relate to the need for finding and managing multiple licences as well as the limited engagement with the actual value creation. For university, there exists the possibility that by encouraging the creation of spin-outs based on university IP, the university engages in higher risk and may lose "star" researchers thus dampening its organisational scientific output. This can prove as an unwillingness for university to support opening data sets.

Commercialisation of public research is a major goal of policy makers. Cervantes and Meissner (2014) explain that weak commercialisation of research can be down to several bottlenecks. They state that information on university inventions is not available enough to potential users, industrial partners' risk and unwillingness to engage with university inventions is compounded by unclear ownership of said inventions as well as different missions leading to misaligned incentives and coordination problems.

A critical challenge in opening up big data is addressed by Katal, Wazid and Goudar (2013). The authors present the potential privacy and security issues related big data collected from user (people) information. They argue the secretive information that a person does not want revealed might come out once their personal information is linked up with data from other sources. Another noteworthy ethical issue with linking up open big data sets is the consequence of using predictive analysis on the newly formed information to identify underprivileged and subsequently treating them worse. The authors also bring up the issue of managing and governing the shared data. The data made available needs to be accurate and complete with standardised API, metadata and formats. The "metadata challenge" is also touched upon by Grabowski and Minor (2017). They imply that simply making data public does not guarantee its usability without the essential

metadata. Metadata as referred by the authors means knowing the identity of the sample and data collection parameters.

Katal, Wazid and Goudar (2013) list multiple technical issues regarding the storage and processing of the amount of data produced. Collecting and linking big data can cause both capacity and performance issue. Even when data is stored and linked correctly the next challenge lies in conducting analysis on the large amount data that can unstructured, semi structured or structured.

Extracting useful information from large open data sets requires expertise and the need for skilled analysts presents its own problems (Hand, 2013). Big data is still a relatively young concept with new technologies emerging requiring new and diverse skill sets. These skills need to be developed in individuals, meaning organisations need to expend resources to introduce training programs and universities curriculum on big data to produce skilled experts in this field for the future (Katal, Wazid and Goudar, 2013). Grabowski and Minor (2017) argue that while progress in storage technologies can help solve the issue of handling huge volumes data, one detrimental challenge for opening big data repositories remains the deep-rooted reluctance of researchers to share their data.

Based on the findings from literature used for this chapter, the opportunities and obstacles for opening big data sets are summarised in table 2 below.

Table 2: Opportunities and challenges in open big data

Opportunities	**Obstacles**
Organisational transparency enables communities to feel empowered	Unwillingness due to perceived risks (time, investment, different objectives)
Reproducible and accelerated research through open access data	Privacy and security issues when combining data sets
New, innovative business models from linking up different data sets	Management, governance and processing of open big data
New products and services based on open data	Researchers reluctance to share due to perceived risks

5 Big data strategies

This chapters lays out approaches that academia use to participate collaborate with industry. Following this, the chapter explores big data strategies, referring to different methods through which big data is commercialised. This includes the assessment of best practices and business models related to big data. Lastly, this chapter present two forms of ecosystem to foster open science practices among researchers.

5.1 Commercialisation of research big data

In the article by Cervantes and Meissner (2014) approaches for commercialising public research are laid out. The authors suggest that patents, licensing income and spin-offs should be used as an indicator to determine the capability of an institution to turn research into innovation. Other channels for commercialising public research are collaborative research partnerships between the public and private sector, staff mobility and contract research and research staff consulting (Perkmann *et al.*, 2013; Cervantes and Meissner, 2014). Universities can also directly exchange knowledge embedded in IP documents with industry to provide access to university inventions on royalty-free and free-free basis. To obtain a higher turnover rate of research-turned-innovations, universities may also encourage their staff to establish new ventures by providing actual incentives, e.g. granting leaves of absence, allow tenure clock stoppage. To promote commercialisation activities from within universities can also consider commercial track record when deciding on staff promotions (Cervantes and Meissner, 2014).

Exploitation routes of university generated intellectual property (IP) is touched upon by Minshall, Seldon and Probert (2007). In their paper the writers explore the pros and cons of two exploitation routes for said IP by either licensing to established firms or creating university spin-out firms. Minshall (2003) explores the conditions under which the option to create a new firm is most appropriate and claims this is true in cases where the technology is platform based and needs substantial investment to further develop.

As remarked by Drexler *et al.* (2014), when trying to combine internal and external R&D efforts organisations face difficulties in identifying external knowledge and see interaction with external knowledge sources as crucial for better innovative performance. To address this, organisations have conducted several activities to reach out to academia. This includes checking scientific publications, journals, analysing university patents and attending conferences and seminars.

An important practitioner of open innovation and commercialisation activities in academia are *entrepreneurial academics*. An entrepreneurial academic is an academic that engages in commercialisation activities that result in patent creations, license sales or new ventures in the form of spin outs. These individuals conduct technology transfer activities in industry collaborations with their goal being more than publishing their research but also the recognition that it has a wider purpose on society. Their involvement with industry may also result in financial benefits and entrepreneurial academics see this involvement as an extension on their research related role that can lead to access to new resources, funding and learning opportunities (Alexander, Miller and Fielding, 2015).

5.2 Big data business models

In their article on commercialisation of big data, Thomas and Leiponen (2016) see big data's value in its secondary use – the so called "data reuse" that enables creation of new products and services. The authors point to the increasing amount of available external open data for organisations to utilise and categorise big data commercialisation under six different models for value creation of big data. The models are listed as *data suppliers, data managers, data custodians, application developers, service providers* and *data aggregators*. The business models generally use freemium, premium and pay-per-use/view as a revenue stream (Thomas and Leiponen, 2016).

Zeleti, Ojo and Curry (2014, p.4) state that open data needs to be "structured, supported, timely, accurate and data releases need to be reliable and sustained over time" to be useful to businesses and for business models to be operational. The authors analyse open data business models found from literature and practice to better define how potential value can be harnessed from open data. In their study they conclude that premium and freemium models, also mentioned by Thomas and Leiponen (2016), are the most used models. The premium model requires customers to pay a premium price for access to the data while the fermium model offers the basic data free of charge but charges a price for more detailed data. The reason for this is its less complicated strategic focus and more successful cases being achieved leading to an increased adoption of these models.

A study by Hartmann *et al.* (2016) examines the data usage of 100 startup companies and the revenue model they are based on. The results point to 73% of startups using external data sources while 76% conduct data analytics as a key activity and 62% rely on a subscription based revenue model. The study (Hartmann *et al.*, 2016) reveals six different

business model cluster types of the startups based on their offering and what they do with the data.

Two types identified as *free data collector and aggregator* and *free data knowledge discovery* collect and aggregate data from different, freely available sources. In these models data is normalised or analysed and is offered to customers with revenue coming from subscription and usage fees. For free data knowledge discovery, there also exists the possibility of relying on revenue from advertising and brokerage fees. The types *analytics as a service* and *data aggregations as a service* analyse and aggregate customer data, respectively. They have similar revenue models and charge based on subscription on usage fees. In the business model labelled as *data generation and analysis*, companies focus solely on generating their own data and may perform analytics on it while getting revenue from asset sales. In the final type, *multi-source data mash-up and analysis*, data from customer as well as external sources are combined, analysed and aggregated. Revenue for these typically comes from subscription fees (Hartmann *et al.*, 2016).

The Figure 1 below contains a summary of the six types of business models and of the key data sources from which data is drawn as well as the key activities conducted with the data.

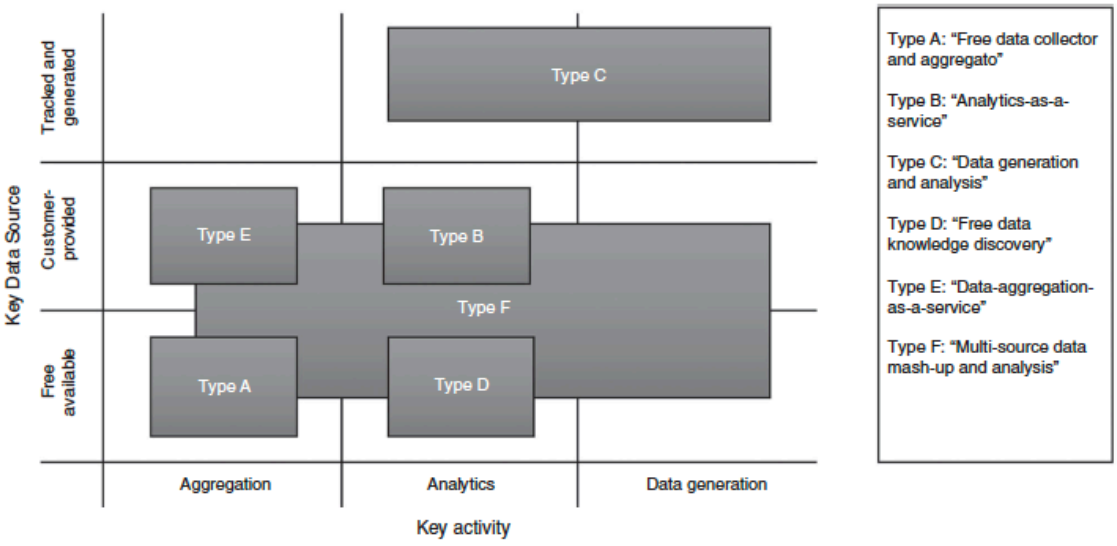

Figure 1: Matrix of data-driven business models (Hartmann et al., 2016)

Researchers looking to commercialise their big data repositories might find most viable use in the models as a data supplier, data generation and analysis or multi-source data mash-up and analysis. As a data supplier, researcher would provide data in the format it is made available in so others may reuse it without putting additional effort into

processing the data. As a revenue stream of freemium or premium model may be easily implanted for these as they are not complex. Using data generation and analysis or multi-source data mash-up and analysis, researchers can use their self-generated data or mix their own data with other datasets and analyse them to provide value while charging based on subscription.

5.3 Big data ecosystems

A remark made by Thomas and Leiponen (2016) is the emergence and crucial role of data ecosystems in making the most value out of big data. This means shifting from "organisation-centric" data to a wider ecosystem where most value is derived from interconnectedness and interdependencies among the data. Such a proposed data ecosystem consists of private organisations, public institutions and end users. Data ecosystems are facilitated by technological platforms, meaning platform owners create standards for the technical system.

The emergence of strategic stakeholders around a technical platform is also mentioned by Ferrando-Llopis, Lopez-Berzosa and Mulligan (2013). They conclude that, at least during the time of their study, there is not enough data from firms to characterise the exact successful business models for such a platform. However, they foresee that the nature of big data and the business models around it causing data ecosystems with technical platforms to appear. Data in this case is produced, owned and handled by different parties resulting in a strong network effect. The authors see organisations developing business models around big data either replicating what industry leaders like Google and Amazon have already done or developing new business models based on technologies that can support data from within and outside.

5.3.1 Virtual research environments

The amount of available open data sets is growing in number government side while publicly funded research projects are also increasingly required to make data openly available. Researchers can use these open data sets in various ways to come up with new data-driven research and generate new datasets, information and knowledge. To make this process easier for researchers a supporting system in the form of virtual research environment (VRE) may be employed. VREs act as an online system enabling collaborative research activities beyond geographical borders and providing researchers

with tools managing complex tasks (Grayling, 2009; Candela, Castelli and Pagano, 2013; Zuiderwijk *et al.*, 2016).

Zuiderwijk *et al.* (2016) note that a big data-driven VRE should have integrated tools for search, accessing, integrating data and fostering collaboration among scientists. The authors propose several requirements for this kind of VRE, among others: data storage, data accessing, data computational services, data curation and data cataloguing. These requirements would allow VRE to provide researchers with integrated open data from different domains and provide open government data in combination with open research data.

5.3.2 Boundary organisations

According to Perkmann and Schildt (2015) boundary organisations can be deployed as an effective tool to facilitate open data collaboration between industry and academia. The authors highlight a case study of a boundary organisation in the Structural Genomics Consortium (SGC) that practiced an open data approach and encouraged innovators to build on the work of others that are deposited in a common data bank. The SGC allows pharma industry partners to disclose their research problems to an audience of innovators from academia by shaping the organisation's research programme. Each pharma company compiles a wish list of proteins they want resolved by scientists. These lists are combined an anonymised into a master list that was never disclosed to the public and circulated for approval board of directors from the sponsor side along with a scientific committee.

Confidentiality is regarded as a key factor and priority for companies as they want to avoid their R&D priorities becoming public knowledge. In addition to appealing to firms, the SGC also pursued strategies to attract and motivate participating scientists. First, they promote the opportunity to work on previously uncharacterised proteins in a state-of-the-art programme. Secondly, the SGC encourages researchers to engage in follow-on research beyond the proteins master list to pursue their scientific curiosity leading to more demanding research and higher scientific impact. This freedom allows scientists to publish high impact articles and facilitated the career progression of participants. The SGC also adopts academic practices by distributing funding to universities so they can employ the researchers on academic terms. The concept of boundary organisation as exemplified through the SGC by Perkmann and Schildt (2015) is working model for open

data collaboration between parties with different interests. This model is visualised in Figure 2 below.

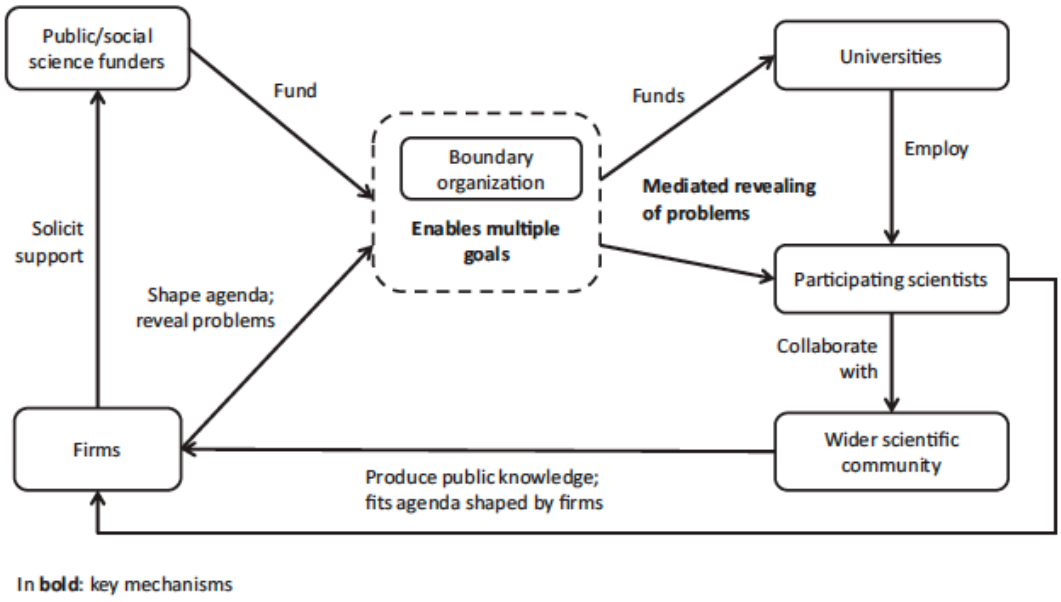

Figure 2: Boundary organisation model for open data partnerships (Perkmann and Schildt, 2015)

6 Conclusions

This thesis presented open big data practices used for value creation along with the opportunities it brings and involved challenges that need to be addressed. The information obtainable from big data through linking data sets and reusing data offer many possibilities for value creation. This value could be realised by making big data open.

While the many benefits of open big data have pushed initiatives and models from an organisational level, it seems that actual open science approaches are still somewhat lacking. This can be attributed to several reasons. First, the scientists' personal motivations, the actual factors that drive their data sharing behaviour, are seemingly not addressed. Their reluctance to share research data stems from their attitude towards data sharing. The factors affecting the attitude are their expected return and worries in data sharing. Secondly, with big data being a recent phenomenon, the challenges that it brings must be addressed to fully enable data sharing. There exist potentially serious ethical and security issues linked with open big data that are of real concern. Technical challenges are related to big data storing, management and processing.

Commercialisation and collaboration models for big data exist. Commercialisation of open government data has seen new business models emerge, some of which could possibly be applied to researcher's big data repositories. Researchers could potentially offer their data as supply, analyse it further for more value or link it up with other data sources to come out with new findings. VREs and boundary organisations offer researchers the opportunity to not only collaborate among themselves but also with government data and industry data, respectively. These two forms of ecosystems could allow scientists see the benefits of open science and open data.

6.1 Suggestion for future research

Open big data ecosystems involving different stakeholders are a good way to enable open innovation in for big data practitioners. There is need to develop guidelines and best practices for data sharing that make data sharing more effortless for researchers while maintaining the data usability for others. Furthermore, how researchers' open big data can be of value to governments, industry and wider society needs additional inspection. These are three key areas worth researching to better comprehend how to make big research data open in the future.

References

Alexander, A. T., Miller, K. and Fielding, S. (2015) 'Open for Business: Universities, Entrepreneurial Academics and Open Innovation', *International Journal of Innovation Management*, 19(6). doi: 10.1142/S1363919615400137.

Alharthi, A., Krotov, V. and Bowman, M. (2017) 'Addressing barriers to big data', *Business Horizons*, 60(3), pp.285-292. doi: 10.1016/j.bushor.2017.01.002.

Boyd, D. and Crawford, K. (2011) 'Six Provocations for Big Data', *Computer*, 123(1), pp. 1–17. doi: 10.2139/ssrn.1926431.

Candela, L., Castelli, D. and Pagano, P. (2013) 'Virtual research environments: an overview and a research agenda', *Data Science Journal*, 12, pp. 75–81. doi: 10.2481/dsj.GRDI-013.

Cervantes, M. and Meissner, D. (2014) 'Commercialising Public Research under the Open Innovation Model: New Trends', *Foresight Russia*, 8(3), pp. 70–81.

Chesbrough, H. W. (2003) 'The Era of Open Innovation', *MIT Sloan Management Review*, pp. 35–42. doi: 10.1371/journal.pone.0015090.

Ciancarini, P., Poggi, F. and Russo, D. (2016) 'Big Data Quality: A Roadmap for Open Data', *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, (June), pp. 210–215. doi: 10.1109/BigDataService.2016.37.

Cowan, D., Alencar, P. and McGarry, F. (2014) 'Perspectives on Open Data: Issues and Opportunties', *2014 IEEE International Conference on Software Science, Technology and Engineering*, pp. 24–33. doi: 10.1109/SWSTE.2014.18.

Drexler, G., Duh, A., Kornherr, A. and Korošak, D. (2014) 'Boosting Open Innovation by Leveraging Big Data', *Open Innovation: New Product Development Essentials from the PDMA*, pp. 299–318. doi: 10.1002/9781118947166.ch11.

Emmanuel, I. and Stanier, C. (2016) 'Defining big data', in *ACM International Conference Proceeding Series*. doi: 10.1145/3010089.3010090.

Etzkowitz, H. and Leydesdorff, L. (1995) 'the Triple Helix---University-Industry-Government Relations: a Laboratory for Knowledge Based Economic Development', *EASST Review*, 14(1), pp. 14–19. Available at: http://ssrn.com/abstract=2480085

[Accessed 20 April 2017].

European Commission (2016) 'The European Cloud Initiative'. Available at: https://ec.europa.eu/digital-single-market/en/european-cloud-initiative [Accessed 3 May 2017]

Ferrando-Llopis, R., Lopez-Berzosa, D. and Mulligan, C. (2013) 'Advancing value creation and value capture in data-intensive contexts', *2013 IEEE International Conference on Big Data*, pp. 5–9. doi: 10.1109/BigData.2013.6691685.

Friesike, S., Widenmayer, B., Gassmann, O. and Schildhauer, T. (2014) 'Opening science: towards an agenda of open science in academia and industry', *Journal of Technology Transfer*, 40(4), pp. 581–601. doi: 10.1007/s10961-014-9375-6.

Grabowski, M. and Minor, W. (2017) 'Sharing Big Data', *IUCrJ*. International Union of Crystallography, 4(1), pp. 3–4. doi: 10.1107/S2052252516020364.

Grayling, I. (2009) 'Achieving success in collaborative research: The role of Virtual research Environments', *Journal of Information Technology in Construction*, 14, pp. 59–69. doi: 11.1039/b000000x.

Hand, D. J. (2013) 'Data, not dogma: Big data, open data, and the opportunities ahead', *Lecture Notes in Computer Science,* 8207, pp. 1–12. doi: 10.1007/978-3-642-41398-8_1.

Hartmann, P. M., Zaki, M., Feldmann, N. and Neely, A. (2016) 'Capturing value from big data - a taxonomy of data-driven business models used by start-up firms', *International Journal of Operations & Production Management*, 36(10), pp. 1382–1406. doi: 10.1108/IJOPM-02-2014-0098.

Hung, D. (2016) 'The Impact of Big Data on Social Media Marketing Strategies'. *Tech.Co*, 22 January. Available at: http://tech.co/impact-big-data-social-media-marketing-strategies-2016-01 [Accessed 3 May 2017]

Jetzek, T., Avital, M. and Bjorn-Andersen, N. (2014) 'Generating Sustainable Value from Open Data in a Sharing Society', *IFIP Advances in Information and Communication Technology*, 429, pp. 62–82. doi: 10.1007/978-3-662-43459-8_13.

Katal, A., Wazid, M. and Goudar, R. H. (2013) 'Big data: Issues, challenges, tools and Good practices', in *2013 6th International Conference on Contemporary Computing,* pp. 404-409. doi: 10.1109/IC3.2013.6612229.

Kim, E. (2014) ‘‘Big Data’ Is One Of The Biggest Buzzwords In Tech That No Ones Has Figured Out Yet’. *Business Insider*, 20 August. Available at: http://www.businessinsider.com/companies-not-embracing-big-data-2014-8?r=US&IR=T&IR=T [Accessed 3 May 2017]

Kim, Y. and Adler, M. (2015) ‘Social scientists’ data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories’, *International Journal of Information Management*. Elsevier Ltd, 35(4), pp. 408–418. doi: 10.1016/j.ijinfomgt.2015.04.007.

Kim, Y. and Zhang, P. (2015) ‘Understanding data sharing behaviors of STEM researchers: The roles of attitudes, norms, and data repositories’, *Library and Information Science Research*. Elsevier Inc., 37(3), pp. 189–200. doi: 10.1016/j.lisr.2015.04.006.

Laney, D. (2001) ‘3D Management: Consulting Data Volume, Velocity and Variety’, *Application Delivery Strategies*, 949. doi: 10.1016/j.infsof.2008.09.005.

Liao, Y. (2015) *Harnessing and boosting university brand in the age of big data*, *Advances in Intelligent Systems and Computing*, 362, pp.497-509. doi: 10.1007/978-3-662-47241-5_42.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H. and Yarkoni, T. (2016) ‘How open science helps researchers succeed’, *eLife*, 5, pp. 1–19. doi: 10.7554/eLife.16800.

McKinsey & Company (2011) ‘Big data: The next frontier for innovation, competition, and productivity’, *McKinsey Global Institute*, (June), p. 156. doi: 10.1080/01443610903114527.

Minshall, T. (2003) ‘Alliance business models for university start-up technology ventures : a resource based perspective’, *11th Annual High Tech Small Firms Conference*, pp. 1–15.

Minshall, T., Seldon, S. and Probert, D. (2007) ‘Commercializing a Disruptive Technology Based Upon University Ip Through Open Innovation: a Case Study of Cambridge Display Technology’, *International Journal of Innovation and Technology Management*, 4(3), pp. 225–239. doi: 10.1142/S0219877007001107.

Okamoto, K. (2017) ‘Introducing Open Government Data’, *The Reference Librarian*.

Routledge, 58(2), pp. 111–123. doi: 10.1080/02763877.2016.1199005.

OpenAIRE (2016) 'What is the Open Research Data Pilot'. Available at: https://www.openaire.eu/opendatapilot [Accessed 3 May 2017]

Padilla-Meléndez, A. and Garrido-Moreno, A. (2012) 'Open innovation in universities What motivates researchers to engage in knowledge transfer exchanges?', *International Journal of Entrepreneurial Behaviour & Research*, 18(4), pp. 417–439. doi: http://dx.doi.org/10.1108/13552551211239474.

Pampel, H. and Dallmeier-Tiessen, S. (2014) 'Open Research Data: From Vision to Practice', *Opening Science*, pp. 213–224. doi: 10.1007/978-3-319-00026-8.

Park, H. W. (2014) 'An interview with Loet Leydesdorff: The past, present, and future of the triple helix in the age of big data', *Scientometrics*, 99(1), pp. 199–202. doi: 10.1007/s11192-013-1123-4.

Perkmann, M. and Schildt, H. (2015) 'Open data partnerships between firms and universities: The role of boundary organizations', *Research Policy*. Elsevier B.V., 44(5), pp. 1133–1143. doi: 10.1016/j.respol.2014.12.006.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broström, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A., Krabel, S., Kitson, M., Llerena, P., Lissoni, F., Salter, A. and Sobrero, M. (2013) 'Academic engagement and commercialisation: A review of the literature on university-industry relations', *Research Policy*. Elsevier B.V., 42(2), pp. 423–442. doi: 10.1016/j.respol.2012.09.007.

Sadiku, M. N. O., Tembely, M. and Musa, S. M. (2016) 'Open Data : Opportunities and Challenges', *Journal of Multidisciplinary Engineering Science and Technology*, 3(11), pp. 6006–6008.

Tacke, O. (2010) 'Open Science 2.0 : How Research and Education Can Benefit from Open Innovation and Web 2.0', *Science*, 76, pp. 37–48. Available at: http://olivertacke.de/wp-content/uploads/2010/11/Tacke-2010-Open_Science_2.0.pdf.

Tenopir, C., Van Der Hoeven, J., Palmer, C. L., Malone, J. and Metzer, L. (2011) 'Sharing data: Practices, barriers, and incentives', *Proceedings of the ASIST Annual Meeting*, 48(1). doi: 10.1002/meet.2011.14504801026.

Thomas, L. D. W. and Leiponen, A. (2016) 'Big data commercialization', *IEEE*

Engineering Management Review, 44(2), pp. 74–90. doi: 10.1109/EMR.2016.2568798.

Trott, P. and Hartmann, D. (2009) 'Why "Open Innovation" is old wine in new bottles', *International Journal of Innovation Management*, 13(4), pp. 715–736. doi: 10.1142/S1363919609002509.

Viseur, R. (2015) 'Open science: Practical issues in open research data', *DATA 2015 - 4th International Conference on Data Management Technologies and Applications, Proceedings*, pp. 201–206. doi: 10.5220/0005558802010206.

Zeleti, F. A., Ojo, A. and Curry, E. (2014) 'Emerging business models for the open data industry: Characterization and analysis', *ACM International Conference Proceeding Series*, pp. 215–226. doi: 10.1145/2612733.2612745.

Zimmermann, H.-D. and Pucihar, A. (2015) 'Open Innovation, Open Data and new Business Models', *Proceedings of IDIMT 2015 - 23rd Interdisciplinary Information and Management Talks*, pp. 1-10. Available at: http://s3.amazonaws.com/academia.edu.documents/38938414/IDIMT_paper_HDZ-AP-final.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1494035156&Signature=e5SfBVFMOp6%2FcdOVab8GqM4s4uA%3D&response-content-disposition=inline%3B%20filename%3DOpen_Innovation_Open_Data_and_new_Busine.pdf.

Zuiderwijk, A., Jeffery, K., Bailo, D. and Yin, Y. (2016) 'Using Open Research Data for Public Policy Making: Opportunities of Virtual Research Environments', *6th International Conference for E-Democracy and Open Government*, pp. 180–187. doi: 10.1109/CeDEM.2016.20.