Designing a Context-aware Decentralized Marketplace for Sensor Data

Master Thesis submitted to the Faculty of Technology, Policy & Management
Delft University of Technology, as part of the Master of Science in Management of Technology

By Raphael Hannaert
Student number: 4632737

Defense on 30th August 2018

Graduation Committee

Chair:	Prof. dr. ir. M. Janssen, Section Information & Communication Technology
First Supervisor:	Dr. ir. A.J. Bram Klievink, Section Organisation & Governance
Second Supervisor:	Dr. H.H. Hansen, Section Energy & Industry
Advisor:	Sélinde van Engelenburg, Section Information & Communication Technology

Abstract

In the past years there has been increasing awareness about the benefits of collecting and using more sensor data for businesses. This has led firms to look for data outside of their boundaries and use some data commercialization mechanisms such as data brokers, and open or privately-owned data marketplace. However, these exchanges solutions are controlled by companies which have a commercial interest that differs from users, leading to lack of transparency and lack of protection of data, loss of data ownership by the provider and no guarantee of fair pricing. These centralized data exchanges call into question the willingness of both data providers and data users to share data. As an alternative, blockchain technology can be used to reduce the control and interference of any firm, leading to a more peer-to-peer and transparent data marketplace. To improve coordination between stakeholders and to enhance a more automated marketplace, the system should be context-aware. The main contribution of this thesis is a proposition of blockchain-based components integrated within a context-aware decentralized data marketplace. Other parts of the system are highlighted, as they need to be subject to more research in order to achieve a fully functional and complete system. Finally, guidelines are suggested for generalization to other types of data and ecosystems.

Executive Summary

Sensor data use has increased considerably in the past years, enabled by a surge in use-cases and better technologies for storing and processing data. Following these trends and the awareness of their potential, businesses have realized the importance of gathering more data than what they actually produce, by collecting data from other sources. Data monetization has therefore also been growing and data has been traded as commodities using several exchange mechanisms such as individual contract arrangements, data brokers, and data marketplaces.

These current exchange mechanisms are inefficient as businesses need to establish partnerships with other businesses individually (contract arrangements), or they present risks and lack of transparency because of the presence of a central company controlling the data flow (data brokers, data marketplaces). This research focuses on efficient ways of sharing data and therefore investigates the second type of exchange mechanisms and their salient features. Data brokers and firm-controlled data exchanges have been criticized for impacting negatively the willingness of both the supply and demand sides to engage in data transactions. The main criticisms involve the lack of transparency, the access to proprietary data by the company, the conflicting interests resulting in opportunistic behaviors, the loss of control by sensor owners over their data, the inefficiency and costs related to unnecessary intermediaries, and the lack of privacy. Following this problem identification, the research question was formulated.

"How can we improve efficient sensor data sharing by reducing risks of opportunistic behaviors in firm-controlled data exchange mechanisms?"

To answer the research question, we investigated the components needed for a design that undermine the firm control in data exchanges, leading to the proposition of a decentralized data marketplace system. This system is based on two main elements, blockchain technology and context-awareness.

Blockchain technology has proven to be an effective tool for removing intermediaries in some processes, as it can force stakeholders to adopt particular behaviors and therefore lead to trust in the ecosystem. More specifically, blockchain characteristics that are used for the marketplace include its immutability and distributed properties, its potential for managing identification and access control to data, and its connection to off-chain data storage via distributed hash tables. In addition, enabling complementary technologies such as smart contracts have been used to propose a decentralized and token-curated data quality check mechanism.

The context-aware system approach is used to deal with the complexity of a decentralized environments. Context-awareness means that the system monitors elements that are part of

the context and react accordingly based on defined rules. In a decentralized data marketplace, there are many stakeholders interacting with each other, with different requirements, frequently uploading and downloading data with various levels of quality. The system must react quickly to these changes otherwise the user experience is affected. In addition, the system requires more automation as there is less presence of a centralized actor managing data exchanges, unlike in the current mechanisms. Therefore, the system needs to be able to sense and adapt to contextual elements. To build this context-aware system, the design method proposed by van Engelenburg et al. (2018) is used.

This method suggests defining first the focus, which is a relationship between objects that need to have a certain value corresponding with the goal of the designer. Then the designer needs to understand which situations lead the focus to take these values. Understanding these situations translates into understanding what the context is. After this context definition phase, the designer must propose sensors and adaptors to interact with this context. The adaptors update the system based on the information collected by the sensors and according to reasoning rules. The understanding of this rules by the computer is made possible since they have been formulated as logic rules using schematic literals.

In this thesis project, the motivations that led to the design of sensors, adaptors and reasoning rules were two situations impacting negatively the willingness to participate in the marketplace. The first challenge is the data quality that needs to satisfy the user's requirements. The second problem is on the provider side and concerns the need for not sharing sensitive data with the wrong parties.

The elements that constitute the context are the data providers and users, the datasets they upload and download, the quality requirements and perception that they have about datasets, and the actual data quality. The actual data quality involves a wide range of features that need to be respected. For instance, quality for a user may be evaluated based on the format, the number of samples, or the data collection method.

The resulting system design consists of some components of a platform supporting data exchanges, including the back-end system (e.g. data manipulation and storage and the blockchain) and the front-end which stakeholders interact with. Basic (i.e. non-context aware) system parts include the blockchain and the connection to the data stored off-chain. This connection is implemented via a pointers system, the distributed hash table. Concerning context-aware components, based on the two problems mentioned above, the following are proposed: Staking tokens on specific data is proposed as a solution for the first challenge, as is an accurate representation of the data quality. Sensors count the number of tokens and convert the result in a quality indicator. If this indicator is superior to the quality required, the adaptor is activated and presents the data to the user; the adaptor is a recommendation system that combines keywords and tokens staked. For the sensitivity challenge, a solution based on blockchain-managed access-control is suggested. The user provides the sensitivity information to the sensors, which leads to an update of the adaptor to grant access solely to allowed businesses. This adaptor is a decentralized permission system using public-private

keys cryptography. The design is described in the form of a Business Process Management & Notation model, summarizing the data uploading and downloading processes.

The current model is not exhaustive as despite having analyzed 10 situations occurring in the context and restricting the willingness of stakeholders to use the marketplace, only the two main ones have been completely exploited for the design phase. The remaining eight others are just considered as high-level guidelines. However, it is necessary to also apply the complete set of steps of the design method to these cases in order to achieve a more complete system, as much context-aware as possible. It is therefore open for future research. In addition, some missing elements of our research and relating to the design were identified: evaluation and comparison of different storage possibilities e.g. cloud vs distributed databases, how to resolve conflicts when there are disagreements about data quality, how to protect data from replication using contracts and/or technology e.g. homomorphic encryption, and economics and business models of the marketplace. Further research should also target other types of data, such as personal data. In this case, the researcher must be careful about the crucial changes that may arise, such as compliance with the General Data Protection Regulation.

Preface

During the past few months, I had the pleasure to work on a topic which combines two of my main technology-related interests: blockchains and data. Despite several changes in the thesis orientation and the limited knowledge about ICT design, having the opportunity to do a design-oriented allowed me to apply what I learned at the faculty of Technology, Policy and Management and during my specialization in Computer Science in Korea. This would not have been possible without the help I was given by my thesis committee, industry experts, and people around me.

I had the chance to have four supervisors who contributed greatly to my research. Marijn Janssen and Bram Klievink guided me in the development of this research, especially bringing clarification about some misconceptions that I had about ICT systems and their relations with intermediaries. Helle Hansen had a significant impact, as she spent much time discussing the structure of my research, and she taught me relevant first order logic concepts. Sélinde van Engelenburg also provided me with precious and continuous feedbacks during these months, and especially she helped me understand and apply the design method she developed. In addition to my supervisors, I would like to thank the TPM blockchain club and more specifically Jolien Ubacht for listening to my thesis presentation and for asking me though questions.

In addition to the help I received at TU Delft, I would also like to thank Chirdeep Singh and the rest of the Ocean Protocol team, Roderik van der Veer, Cassandre Vandeputte, Abe Scholte and Brian Manusama, as well as Harm van den Brink, for providing me with insightful information through interviews.

Finally, my family, colleagues and friends at Bitcoin Center Korea, and my friend Sam Sadraee provided me with an important support during these past months. Thank you for your patience and feedbacks.

Page intentionally left blank.

Table of Contents

List of figures

List of tables

Chapter 1: Introduction

1.1 Data as commodities

Data is employed for a very broad range of applications across all industries and bring advantages to many stakeholders (Zuiderwijk et al., 2014). It helps decision-makers to take data-driven decisions or increase security (O'neil, 2016), researchers to study phenomena, businesses to better know their customers, fit and develop strategies (Chen et al., 2012), among many other purposes. In particular, it serves artificial intelligence applications where machine learning capabilities are able to manipulate and extract insights from very large data sets, leading to increased accuracy and applications (Halevy et al., 2009). With the development of information technologies, considerable amount of data is now produced and stored for direct or potential future usage. The quantity of data usage keeps increasing, especially with the emergence of new data-intensive (i.e. that produce significant amounts of data) applications such as the internet of things which is forecast to constitute a $4 to $11 trillion economic impact by 2025 (McKinsey report, 2015).

1.2 The information silo problem

Despite the need for data and its increasing supply, most data remain unused (interview 3, appendix A). It is just stored in local databases or on other storage solutions such as clouds, which are provided by businesses. These hosts store huge amounts of data, as illustrated by the size of some of their data centers. This implies that there is a potential that is not fulfilled as other parties that could benefit from data are not able to use it. This is referred to as the *information silo* problem and results in stakeholders carrying out redundant work by having to look for information and build datasets, which could be available from other parties.

1.2.1 Illustration: developing and commercializing autonomous vehicles

The development of autonomous cars (i.e. self-driving cars) is a typical case to illustrate the importance of data sharing, as it is one of the numerous fields requiring a significant amount of data for various purposes.

In the first place, pattern recognition algorithms need millions of data entries to be able to understand the environment. As an example, an image recognition algorithm using machine learning needs millions of labelled pictures in order to be able to recognize elements such as roads, cars, pedestrians, or traffic lights. Labelling pictures manually requires more labor resources than a single firm is able or willing to invest. Instead of replicating the work within each firm, having the possibility to acquire the labelled pictures from an external source allows them to save the effort. By having an entity, external or one of the firms, carry out the labelling and making the data available, the companies can avoid redundant data creation.

Image recognition idea is only one example. Self-driving cars include many other sensors and algorithms that need to be trained with other data.

A second case where autonomous cars would have to combine their data is in order to prove the safety of self-driving vehicles with a sufficient level of certainty. Kalra (2016) has demonstrated that 275 million miles are required to have enough data to demonstrate that autonomous vehicles cause less than 1.09 fatalities per 100 million miles driven with a 95% confidence rate (i.e. proving that autonomous cars are safer than human-driven cars based on the fatality rate). Proving such facts is the type of request that could be required by regulatory bodies before allowing such vehicles to be commercialized. However, it would take more than 12 years to demonstrate this fact even with 100 self-driving cars in the fleet driving permanently, which is more than any company currently has (Kalra, 2016). This illustrates the difficulties that data-driven algorithm developers are facing as the collection of so many data points is a cumbersome task. In particular, such numbers are unreachable for most startups or companies with less resources and therefore it stops them from participating in the development, even if they have the best machine learning practitioners and algorithms.

1.2.2 Current data exchange mechanisms

Over the years organizations have come to realize this potential created by acquiring datasets beyond the ones they produce internally (Gopalkrishnan et al., 2013), and some solutions for accessing external data have been proposed and implemented, as an attempt to solve the information silo problem. In addition to one-to-one business contracts as data exchange agreements, such solutions include buying information from data brokers which are generally specialized entities that collect data in a particular or several fields and sell these with a commission or some monetized added-value (Federal Trade Commission, 2014). Major platforms have also been developed in order to share data, like open data marketplaces generally supported by the public sector (Zuiderwijk, 2014), or other marketplaces owned by companies such as Microsoft Azure Marketplace or Infochimps. These marketplaces provide the information on-demand at any time and from anywhere (Truong, 2011).

1.3 Problem statement

Section 1.2.1 has highlighted the need for data exchange mechanisms to cope with the increases in supply and demand for data. Section 1.2.2 has highlighted four current main possibilities for sharing data: data contracts, data brokers, open data marketplaces, and privately-owned data marketplaces. However, these current solutions show major disadvantages. Data contracts have a very low efficiency as each actor needs to enter in contact directly with another one and reach an agreement. Open data marketplaces are limited in terms of pricing models (i.e. data are normally free) and usage restrictions (Janssen et al., 2012). In addition, they focus more on government to business exchanges.

The two last options, data brokers and privately-owned data marketplaces, are more effective for exchanging sensor data because businesses do not need contact with each data provider directly and because opportunities for increasing revenues make it more attractive for data providers. They also do not require for each transaction a specific contract that would have to be signed in person or via traditional communication tools. Nevertheless, they present a major drawback: there is a company acting as a gateway between the data supplier and the data user, and this company has a commercial interest that differs from other stakeholders. In other words, a company that controls a data marketplace or some parts of an exchange process (e.g. storage) is granted large power as it can benefit from interferences in the exchange process. Their interests may conflict with other stakeholders, leading the firm to take malicious actions towards users. Interference may involve data storage, payment gateway, data flow infrastructures and the related benefits depend on the extent to which the company has control over the exchange process. For instance, as a data storage provider, the firm has the ability to decide upon whom to share the information with or getting insights from unencrypted data (despite being bound by a legal agreement). The lack of transparency of these companies has been criticized e.g. data broker's practices has raised questions about their trustworthiness (Federal Trade Commission, 2014). The direct consequence of this risk is the unwillingness to participate in data exchanges by data providers and users, as the perceived risks may outweigh the benefits (Interview 1 appendix A; Roman & Stefano, 2016).

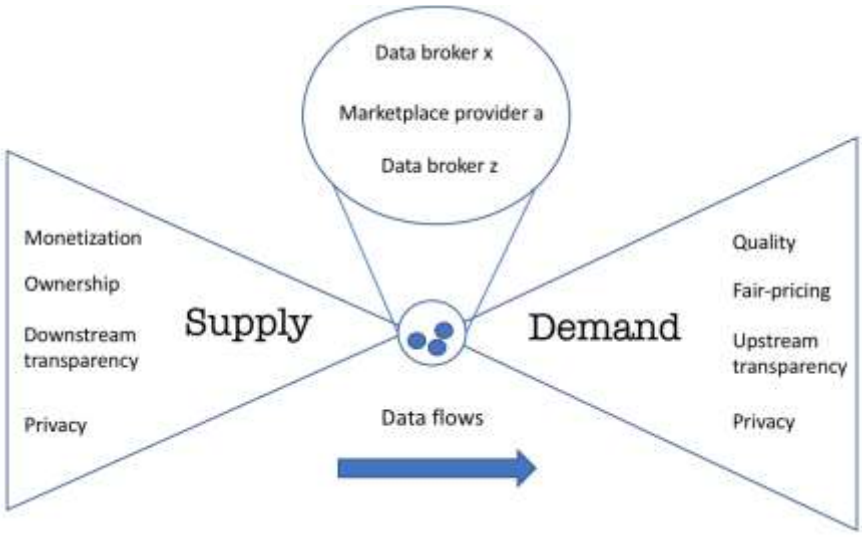

Figure 1: Current firm-controlled data exchange mechanisms

This thesis aims at designing a new system that enables direct transactions of data between sellers and buyers, with limited intervention opportunities by the third-party companies owning or building data marketplaces. By limiting these opportunities, we aim at resolving the existing problem of trust.

In addition, the marketplace must be aware of the context to provide users with the necessary level of experience and enable more automation. This is justified by the very complex environment, involving various stakeholders with requirements that may differ, and elements constantly evolving, such as the datasets uploaded on the platform. To cope with this complexity, the marketplace must sense the relevant parts, and constantly updates its behavior e.g. present the right datasets to users via a recommendation system, as will be explained in Section 3.3.2. More automation is required to reduce the required trust in the company, by increasing transparency. The code governing the actions of the platform can be open-sourced, and the algorithms may manipulate datasets (e.g. presenting to users) without making these available to the firms building the marketplace. This is also further described in Section 3.3.2.

The system should also meet basic requirements that translate values to be considered from both sides of the marketplace. Requirements for the data provider side include data monetization, conservation of data ownership, downstream transparency (e.g. who is using the data? For which purpose?), and privacy. For the data user side, quality of data, fair-pricing, upstream transparency (e.g. data collection methods) and privacy are relevant characteristics which will be discussed. Finally, one may argue that individual data contracts as mentioned above are already a peer-to-peer process. It is true; however, the new system also needs to be efficient by creating many-to-many exchanges and therefore a data marketplace platform architecture will be at the core of the system. Figure 1 summarizes the problem statement, by representing the firms controlling the data exchange as the only gate for massive (and therefore not taking into account the direct individual contract between two businesses) data exchanges between supply and demand. It also illustrates the requirements of both sides.

This thesis will focus only on sensor data for businesses, i.e. data provided by sensor owners to businesses. These data are proprietary and can lead to revenues for the providers, but they are not personal. We believe that this scoping choice is relevant for several reasons. First of all, as mentioned in the introduction, sensor data production is growing fast and will continue to do so in a more automated way, as the internet of things is rising. Second, personal data are subject to strong protection regulations, which add much complexity to an already complex problem. Finally, sensor data measure physical phenomena, which are possible to measure by any agent equipped with the right and functional sensors. This can help judging data quality since several measurement of a same factor can be taken by several parties and compared. For building a decentralized data marketplace, this also decreases the complexity.

The removal of the central firm is represented in Figure 2, which if compared with Figure 1 indicates a transition from current centralized exchanges to a more peer-to-peer system i.e. with a direct data flow from supply to demand. This is further articulated by the quote from Karafiloski and Mishev (2017).

"By removing the central authority out of the system, there is no longer a mediator processing the actions and the data. That results with lower transactional costs, non-reversible transactions and no need for trust in the governments or private corporations." – Karafiloski & Mishev (2017, p.763)

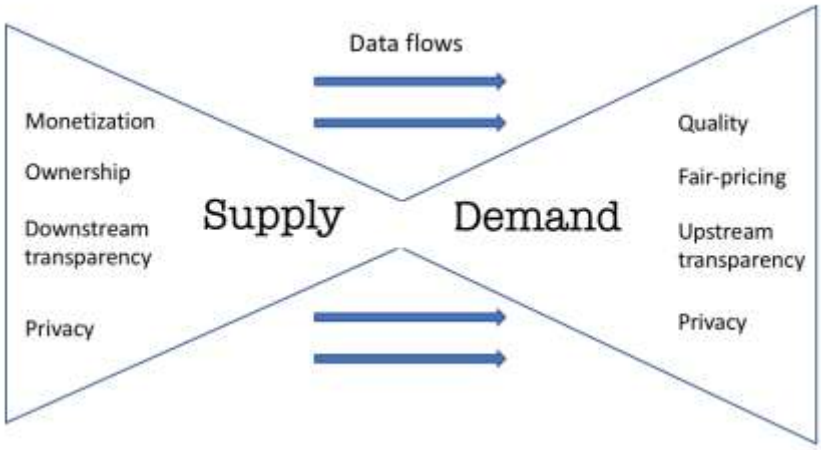

Figure 2: Decentralized data marketplaces: direct exchanges

In the literature it has been suggested that a new architecture for sharing data is necessary in the current "data lake" era, referring to the abundance of data produced (Khine et al., 2018). Several interviews with industry experts have confirmed this need (Interview 1, 2, 4, appendix A). Roman & Stefano (2016) claim that in order to have organizations participate, there is a need for a common trusted data marketplace and ask for research about guidelines and standards for this environment. They also mention that not only technology, but a legal framework and policies are required to reach the security that can enable this trust. They suggested further steps in research namely identifying concrete technologies to work with and integrating these enabling technologies in a data marketplace system.

1.4 Approach

There is therefore an issue to be solved: how can we limit the opportunistic behaviors of firms controlling data marketplaces. This thesis proposes to resolve it by the means of a design. However, the current scientific knowledge does not include what such a design should look like. The design-oriented research output is to define this system in order to enable actors to develop and implement solutions based on this system. Two main elements are important for the design: blockchain technology and context-awareness.

1.4.1 Blockchain-based solution

Blockchain technology has been illustrated as an potential effective artifact to achieve more decentralization in digital asset exchanges (Swan, 2015; Tapscott & Tapscott, 2016). Decentralization represents the removal of intermediaries in processes. Complete removals are difficult to achieve as blockchains as a technology also bring new types of intermediaries (e.g. the developers of the blockchain-based platform), it is more accurate to talk about reducing omnipresence and opportunistic behaviors of companies owning marketplaces. Blockchain technologies offer the necessary tools to force or at least incentivize stakeholders to act in specific ways, and therefore offer a possibility to reduce these opportunistic behaviors.

The first instance of a protocol using blockchain technology is Bitcoin (Nakamoto, 2008), as a way to transfer money directly between individuals (peer-to-peer). In addition to financial applications, blockchain technology has also been used for other *decentralized applications* such as to implement peer-to-peer distributed file storage system (Benet, 2014; Buterin 2014), decentralized prediction markets (Peterson, 2015) or for creating decentralized autonomous organizations (Tapscott, 2016). By analogy, blockchain will be used in this design research as a technology to reduce risks of firm-controlled marketplaces, forming what will be called *decentralized data marketplaces*.

1.4.2 Context-aware marketplace

Context-awareness refers to the ability of systems to adapt their operations to the current context without explicit user intervention. It has been vastly used in mobile applications (Schilit and Theimer, 1994) and/or for recommendation systems (Baldauf, 2007).

Context-aware systems can sense and adapt to the relevant part of their environment, which is a crucial property that our system will need to demonstrate. This need can be articulated based on the following reasons: first, the data marketplace environment is characterized by many factors evolving dynamically, such as stakeholders and their requirements about data sharing. This is emphasized by the decentralized nature of the marketplace. Decentralization makes the system more complex (Buterin, 2016) i.e. there is no single party to coordinate the actions, to take responsibilities, to decide the rules. In these environments, it is difficult to target which elements are relevant and how should the different parts (stakeholders, actions) connect with each other e.g. determining the dependencies is not done centrally anymore. Secondly, matching supply and demand is the core of a marketplace and should be done effectively with some kinds of data recommendation system; and previous works indicates that context-awareness is highly relevant for recommendation systems (Baldauf, 2007). The marketplace also requires a higher degree of automation since we aim at removing the human control over the marketplace. In addition, customer experience is important to improve the

willingness to participate in the marketplace. We argue in 3.3.2 that these three features are achieved more effectively if the system can sense and adapt to context.

To build the context-aware marketplace, we will use a design method that was proposed by van Engelenburg, Janssen and Klievink in their paper "Designing context-aware systems: a structured method for understanding and analyzing context" (2018). This choice is justified by the exhaustive description targeting designer, and by the available support from the authors.

1.5 Thesis structure

Following this introduction, Chapter 2 describes the research question and sub-questions, as well as the research objective, validity and relevance considerations. Chapter 3 states the research approach that will be used to answer these questions, including the literature review and interviews to form the knowledge base, as well as the description and relevance of the context-aware design method. The knowledge base upon which new scientific knowledge will be built is described in Chapters 4 and 5. Chapter 4 gives a description of the importance of data sharing and especially the current ways to exchange it, and Chapter 5 offers a basic understanding of blockchain technology in terms of adoption, values, and technical architecture, as well as why it is relevant for our case. Then components that will be used in the design phase are introduced. The design method is then applied to construct a sensor data marketplace in Chapter 6, which is the main chapter of this thesis in terms of knowledge creation. Chapter 7 concludes by answering the research questions, and Chapter 8 provides the reader with suggestions for generalizing the model and reflections about data marketplace adoption, blockchain diffusion, and the context-aware method used.

Chapter 2: Research questions

2.1 Introduction and objectives

In the problem statement section, the problem has been described at a high-level: exchange means that are governed by a single actor or group of actors present some risks and require a certain level of trust in these stakeholders. As a consequence, it discourages some businesses from sharing their valuable data. The introduction has also briefly introduced blockchain technology as a way to reach more decentralization. The goal of this research is:

"To understand and resolve the problems related to trust in firms controlling sensor data exchanges, including centralization of important parts of data exchange processes. As much as possible, other qualities and requirements of existing data exchange methods should be preserved. By 'resolving' we mean suggesting a new decentralized data marketplace system and describing how it should interact with its context. Finally, guidelines should be delivered about how to generalize the system to other cases."

2.2 Research questions

To reach the aforementioned goal, we formulate the following main research question:

"How can we improve efficient sensor data sharing by reducing risks of opportunistic behaviors in firm-controlled data exchange mechanisms?"

Note that the main research question includes the word "efficient". This implies that the simple mechanism of having only one-to-one contracts between businesses is not considered for the literature review nor for the design phase, as this method has been considered as inefficient by the author since it is evident that this practice is not scalable. The lack of scalability results from the fact that each business needs to enter in contact with sensor owners (or the other way around). The focus in this research is about many-to-many data exchanges.

To define the sub-questions, we go back to the problem statement and extract the main elements that need to be investigated.

First, *understanding* what the current solutions to exchange data are is crucial as there is no sense in reinventing existing mechanisms. Basic literature review has led us to identify some ways to share data, but an exhaustive description is required. This leads us to the first sub research question, where "solution" is to be understood as not only the technical infrastructure but also the environment surrounding it. More specifically, this part suggests

understanding what the data exchange ecosystem is, who are the stakeholders, their values, and how do they interact. This part can further be decomposed in the various data exchange mechanisms in order to be able to analyze more extensively each mechanism.

Sub-question 1: What are current scalable solutions used for data sharing between sensors owners and data users?

After answering this first sub-question, the related problems that stakeholders may perceive with a centralized data marketplace propositions are described. In particular, the trust problem in firm-controlled data exchanges needs to be investigated as it is this criticism that has been emphasized by developers of decentralized data marketplaces (Ocean Protocol Foundation, 2017). The trust problem is also a broad term that will be detailed in the section 3.1. Formulating the problem in terms of trust in equivalent to the formulation of the main research question which mentioned opportunistic behaviors. A process which relies more on trust in people instead of control gives more opportunities for these people to behave unexpectedly (Tapscott & Tapscott, 2016)

Sub-question 2: Why is the trust in a company controlling the data exchange a problem?

Once the problem of trust has been established and decomposed into its main components, we start resolving it. Since we want to build a context-aware system, we also need to understand what should be part of the context of the marketplace we want to build.

Sub-question 3: Which parts of the environment belong to context?

For the design phase, we focus on the use of blockchain technology. The sub-question 4 written below serves to present the parts of blockchains, from a theoretical point of view, that will be required for the design of the system. It is important to note that blockchain per se is a broad topic, and that we provide only the relevant information for the sensor data marketplace design.

Sub-question 4: Which blockchain applications or properties can be used to achieve more decentralization efficiently in a sensor data marketplace?

Based on the four sub-questions, we can answer the main research question aforementioned.

Finally, it is necessary to assess the marketplace design, including looking at what the marketplace can do and if it is exhaustive. If not, what should be added, and how does the marketplace fit with other cases. Not only the outcome should be evaluated, but also the design method and more specifically important choices that have been made.

Sub-question 5: How can the designed marketplace be evaluated?

In Figure 3, a diagrammatic overview of the research questions is presented, highlighting the different phases of the research. After building the knowledge base by answering question 1 to 3 and describing relevant blockchain properties, the design method can be executed to propose a context-aware blockchain-based decentralized data marketplace, using sub-question 4. The design is then evaluated (sub-question 5). As illustrated on Figure 3, the resulting system is the answer to the main research question and therefore constitutes the thesis outcome. In the discussion part, guidelines should be proposed to improve external validity i.e. generalization. In this discussion part, future research opportunities are also suggested.

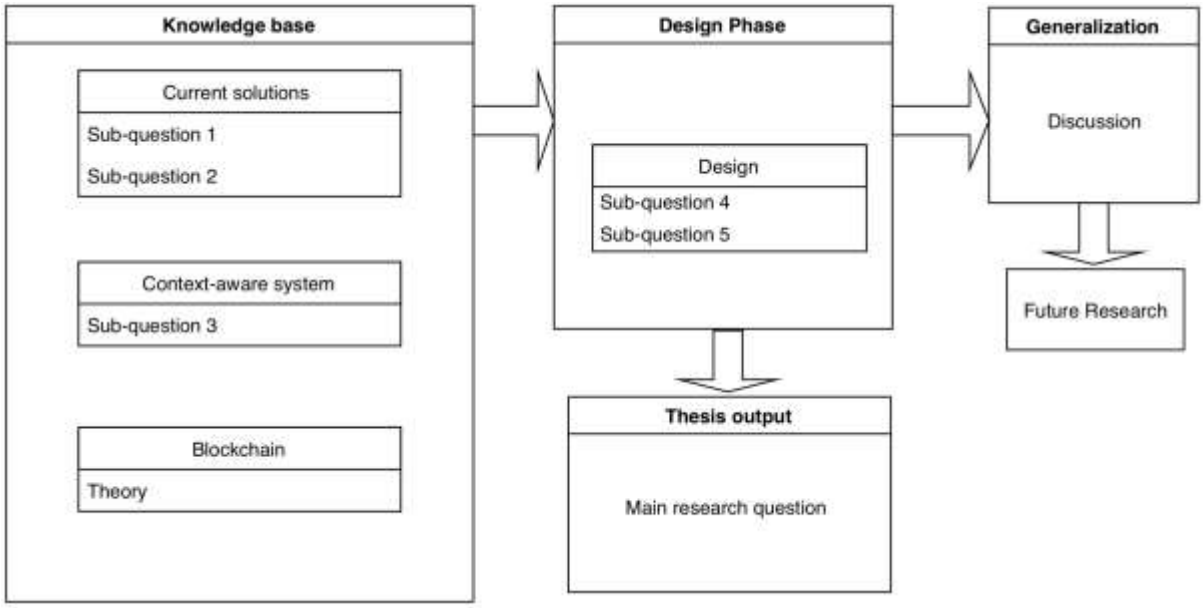

Figure 3: Research questions plan and main phases

2.3 Research outcomes

2.3.1 Scientific contribution

The expected output of this thesis is a context-aware system for sharing sensor data between sensor owners and businesses, using blockchain technology. Therefore, this research creates new knowledge by combining a new technology (blockchain) and a new context-aware design method, with an emerging need (sharing data) and its related problem (centralization).

2.3.2 Scope

An exhaustive design of the system is not in the scope of this research since the resources are limited. However, once components have been proposed and the overall decentralized data exchange process is described, it is possible to pinpoint which elements can be subject to further research. Therefore, in addition to contributing to science with the design, we contribute to science by giving suggestions about how to complete and extend this model. This will be done in the discussion section. As a complement to this main scientific contribution, considerations will be added in the Section 8.1 for generalizing the design to new cases. New cases may include different stakeholders other than businesses (e.g. individuals, governments), different data types such as more personal data which are subject to stronger rules like the General Data Protection Regulation. The considerations will motivate further research in this emergent field.

"Sensor owners" has a broad meaning as it includes agents who own sensors per se, but also if they possess devices composed of sensors which primary role (of the device) is not the measurement in itself (e.g. a mobile phone aims at making calls but is equipped with sensors for various purposes). Discussing the ownership of data outside of the marketplace is out of the scope as this thesis focuses on the exchange process between a data owner and a data user. The assumption is therefore that data providers owns the sensor and all data produced by these sensors. In addition to being proprietary (i.e. belonging to their owner), the sensor data to be exchanged on the marketplace is not personal. Elements such as the General Data Protection Regulation are therefore not part of the context.

2.3.3 Societal and managerial relevance

This research focuses on the exploration of a technology (blockchain) and how to leverage it to build a product (context-aware data marketplace) that for instance businesses can use to improve their performance while maintaining privacy and sensor owners can keep control over their data, justifying their managerial relevance and therefore justifying the relevance of this thesis for the Master of Science in Management of Technology. With this data marketplace actors can save considerable amounts of time and effort, as well as increase their innovation abilities by having access to data they could not have without it. Digital transformation is now a major part of the agenda in both public and private sectors, and actors have more awareness of the potential of data, as well as how to get insights from data using analytics or statistical methods. This research is therefore part of a major societal phenomenon which currently does not contain the right approach to enhance massive data exchanges as there is a mismatch between current unreliable data exchange solutions and the high value of some information.

In addition, this research aims at providing entrepreneurs and platform developers with the knowledge base required to build a decentralized data market that can be used by firms and other stakeholders. Based on the findings of this research, they will not only know why there is such a need for decentralization of data exchanges, but also what are the relevant elements to implement in the platform, and how the elements should be aligned. As a consequence, this thesis also has a practical component (due to its design nature) as it contributes to the knowledge required for developing data marketplaces. If these data exchange platforms become vastly used it would impact positively and significantly data sharing.

By exploring a cross-industry managerial challenge and suggesting a concrete solution using technology as a corporate resource that will help firms overcome this challenge, this research has a direct relevance for the Master of Science in Management of Technology at the faculty of Technology, Policy and Management from Delft university of Technology.

Chapter 3: Research approach

This section introduces the research approach used to answer the research questions. Section 3.1 presents the outline of the research, with the inputs, outputs and dependencies required at each stage. Section 3.2 describes more in depth the knowledge base formation part with the literature review and the interviews methods, and section 3.3 describes the approach for the design phase.

3.1 Outline

The research methods are divided into two parts coinciding with the knowledge base creation and the design phase. For the former, information is collected using two methods: a literature review and semi-structured interviews. The transcripts of the interviews are then analyzed to extract insights. For the design phase, we use the aforementioned context-aware design method.

3.1.1 Knowledge base

To build the required knowledge base upon which the new knowledge can be created, the main research method is a literature review. There are five main outputs necessary to extract from the literature, based on the four research sub-questions: (1) Current data exchange mechanisms, (2) the problems associated with the centralized control by a company, (3) relevant blockchain applications to solve these problems, (4) context-awareness, including reviewing the design method and (5) the current state of decentralized data marketplace research. In addition, some more information on problems of current design mechanisms and about the context is gathered using interviews. The outputs of (1)-(5) will serve as direct inputs for the next phase as the knowledge will be applied directly. The "current design mechanisms" will not serve directly for the design phase.

3.1.2 Design phase

With the outputs from the knowledge base part, we will proceed to the second part: the design phase. At this point, the problems that we are solving, and the context of the system will be clear, as well as the tools to make our design: the necessary technology knowledge (mainly blockchain-related) and a design method. The method is then applied and results in the elaboration of components to be integrated in the architecture, as well as rules about how they should behave according to the context. One could argue that the context is part of the design method, as proposed by the paper we use. However, we consider this still in the knowledge base since it constitutes information based upon which we will build the actual

design. To some extent, it can also be considered as a bridge between the literature review and the design phase.

3.1.3 Generalization

After the design is proposed, it is important to validate it against other cases, and to identify the changes that need to be implemented for improving external validity. Suggestions for future research about applying the model to other sectors will be proposed.

3.1.4 Research diagram

The research diagram below on Figure 4 summarizes the different phases already introduced in Chapter 2 and replace these sub-questions by the research methods used to answer to the questions, with the other blocks being the outputs expected from each research step. First of all, we use a literature review to get information about the current data exchange solutions and their critics, the design method that we will be using, blockchains, what belongs to context (since we are building a context-aware system). Further information is collected about the problems of current data exchange mechanisms and the context using interviews. The answers from the interviews must be analyzed. The findings are used for the design part, which results in the various components and rules that are integrated to form the thesis output: the decentralized data marketplace. Finally, results from the design phase and literature review are used for providing generalization suggestions. More specifically, we will suggest in Section 8.1 to use analytical generalization to determine guidelines. We will provide more details for this future research.

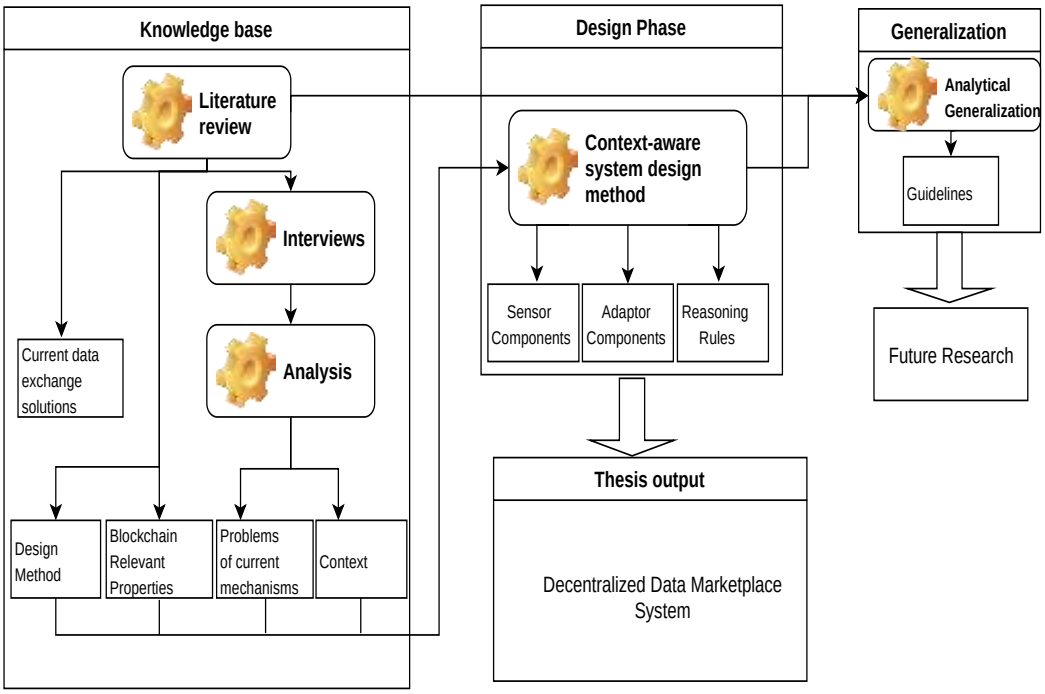

Figure 4: Research approach diagram

3.2 Constructing the knowledge base

3.2.1 Literature review

The first method used in the thesis is the systematic literature review. The systematic literature review aims at providing transparency about the choices made by the researchers when they search for information, take decisions or conclude. The objective is to keep the research scientific by removing the bias that the authors may have (Tranfield, 2003). This review methodology aims at constructing a solid knowledge base upon which our design phase will be based. The literature review must be completed with a strategic approach in order to maximize efficiency and ensure that the relevant papers can be found. The first element of this strategic approach is to divide the research into five main domains investigated which are as much as possible mutually exclusive and collectively exhaustive. In other words, each category should focus on a different topic, but together all areas should cover the entirety of the required knowledge base.

According to the research outline, this section will explain the literature review methodology for the six areas mentioned in 3.1.1. As a reminder, these are: (1) Current data exchange mechanisms, (2) the problems associated with the centralized control by a company, (3) relevant blockchain applications to solve these problems, (4) context-awareness, including reviewing the design method and (5) the current state of decentralized data marketplace research. The findings of the literature reviews will be crucial for the following parts about

the design. In addition, there is a sixth topic which was done as the first step of this research in order to establish the managerial relevance: (6) research about the importance of data.

Concerning the methodology, we now elaborate on what databases will be explored to find relevant articles, and with which search criteria. This will depend on the specific area of research, as some topics may be more technical than others. For instance, on the one hand a hypothesis is that blockchain-related papers can mainly be found in computer science databases for journals such as the digital library IEEE (Institute of Electrical and Electronics Engineers) Xplore. On the other hand, understanding what belongs to context may be found mainly in databases that contain more societal and information management content, such as Scopus. Despite the vertical classifications of papers per domain, we used first cross-domains search engines in order to explore directly different databases by looking for specific metadata. The search engine used for this research is Google Scholar. Altogether, 39 papers, articles, or other documents (e.g. grey literature including articles and whitepapers) have been collected for the literature review phase. The figure below illustrates the number of documents found and exploited per category. Finally, as data usage and exchanges have a consequent managerial component, reports have also been collected from main management consultancy firms.

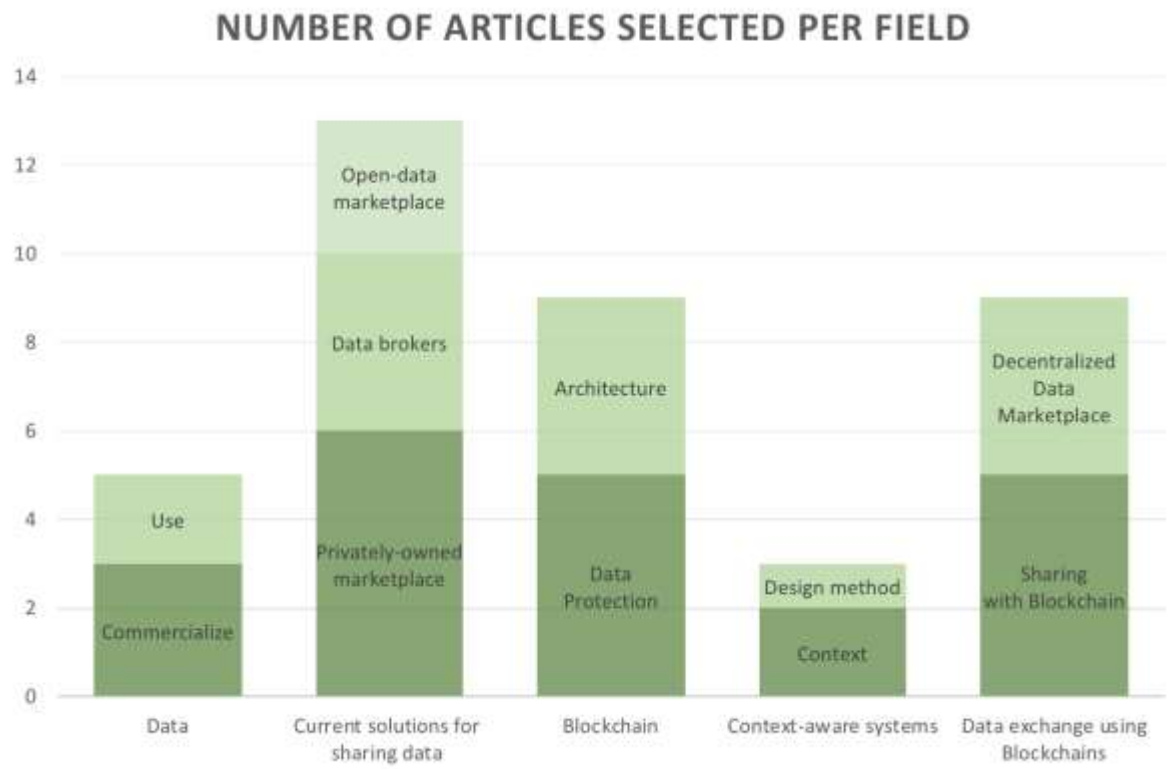

Figure 5: Articles selected per research focus

On Figure 5, we can see on the bar chart the main areas to explore with the literature review and the number of articles selected per area. More specifically, the areas are divided into different topics. Data importance is split in importance of use of data, and of potential related to commercialize data. Current solutions are divided into the three main ways to share data,

namely via data brokers, privately-owned marketplace, and open-data marketplace. The blockchain area is broad, but only the relevant topics are selected: (1) architecture. As we are building a blockchain-based system, it is important to understand how the technology works and what elements should be considered in the design. (2) blockchain applied for data protection. We justify this second topic by the fact that blockchains and data are the main foci of this thesis, and data sharing methods are only suitable if data can be protected. Otherwise actors could just acquire data and make it available without the need for marketplaces. The fourth area is about the notion of context-awareness, divided in a theoretical part about context-awareness (i.e. providing background definition and understanding) and a design methodology part. The former brings an understanding of what context is and why is it important, while with the latter we are looking for a method to design such a context-aware system for sharing data. The last part is about the current literature in using blockchain for sharing data, or even more specifically on decentralized data marketplaces directly. Finally, problems associated with the centralized control by a company do not have a specific bar in the chart despite being mentioned as one of the six areas. These are discussed across the other areas, mainly in articles about data mechanisms and about decentralized data marketplaces, since these solutions are usually proposed to target the current flaws which are therefore highlighted.

Importance of data

The first literature that is reviewed is about the societal and managerial relevance of data. The databases mentioned above were explored, by using keywords “data AND importance”; “data AND commercialization”; “data AND application”; “Data AND business”. Following this search, five papers have been collected, two about the use of data and three about the commercialization potential. One of the papers (Karafiloski et al., 2017) in the commercialization of data is a review of the existing literature about the topics and was therefore the first paper analyzed to have a comprehensive view of the literature state in this field. There were definitely many more relevant papers that could have been selected for this part. However, as it is only an introduction to our topic, getting the insights from the literature review paper as well as some complementary information in 4 other papers selected based on the number of citations (e.g. 3031 citations for Chen et al. (2012) and 806 citations for Norvig et al. (2009)) has been judged satisfying to understand the current state of data use and the need for data exchanges.

Current data exchange mechanisms

This is a highly multidisciplinary field, as illustrated by Figure 6. This figure has been obtained by searching for the combination of terms “(data OR information) AND (exchange OR sharing OR marketplace)" on Scopus. There are 366,000 documents results based on this search. Especially, 9.2% of the articles come from business management (3.6%) and social sciences (5.6%) research. This shows that there is a large number of documents available in

our research field (about 35,000), which demonstrates the relevance of data sharing for business and social science applications. In addition, sensor data can also be classified as part of engineering and computer science because of the inherent nature of sensors. As a consequence, our research is also very multidisciplinary.

Figure 7 represents the evolution of research in the field, with a continuous increase since 1995. We have been able to identify the relevant mechanisms by filtering the results to the social sciences and business management part, and successively with the computer science and engineering part. The latter choice of filtering in more technical fields is justified by the design side of this research, as it involves an ICT architecture within the system and therefore insights from engineering and computer science may be relevant.

13 articles have been selected based on their topic (i.e. we tried to have at least 3 articles about each type of data exchange), on the number of citations, on the ability to enter easily in contact with the authors to ask clarifications and complementary information, and on the importance of the source (e.g. articles coming from the US Federal Trade Commission). These 13 articles are therefore (first argument of this paragraph) further divided into three categories corresponding to the different methods used: privately-owned data marketplaces (6 articles), data brokers (4 articles), and open-data marketplaces (3 articles).

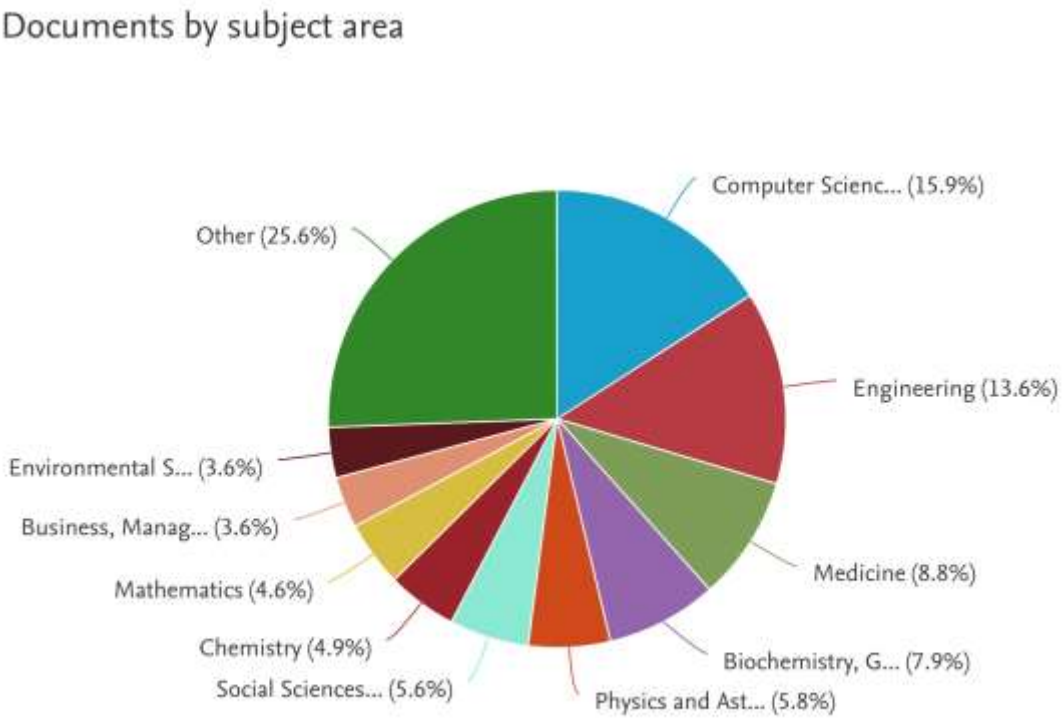

Figure 6: Documents by subject area, using keywords "(Data OR Information) AND (exchange OR sharing OR marketplace)". Retrieved from Scopus.

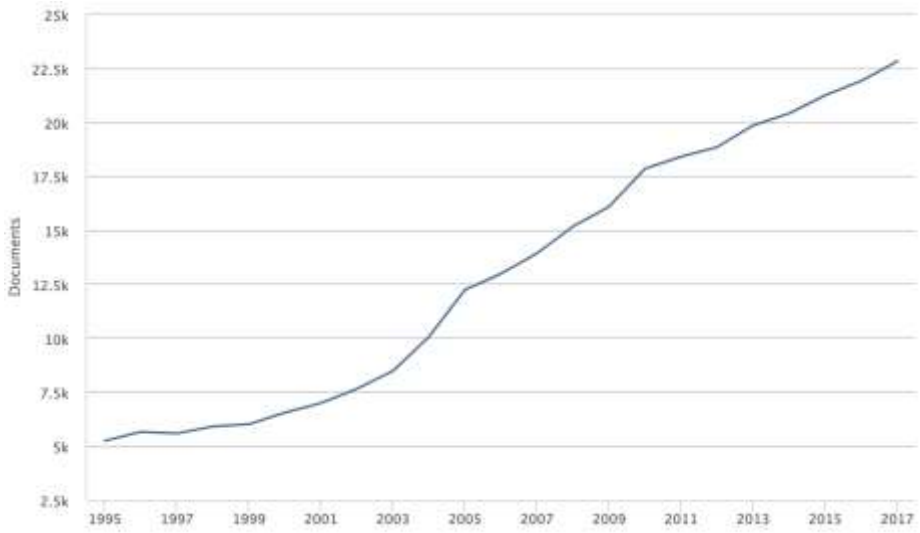

Figure 7: Scientific documents over time, using keywords "(Data OR Information) AND (exchange OR sharing OR marketplace)". Retrieved from Scopus.

Relevant blockchain properties

According to Figure 9 below, blockchain technologies are an emerging topic in research as the number of articles produced has surged in the past year. Blockchain research is mainly conducted in computer science and engineering which is unsurprising regarding the information technology nature of blockchains. Nevertheless, there are also hundreds of documents about blockchains in social sciences, business, economics, and decision science, according to Figure 8. In the literature it is expected to find descriptions of blockchain technologies, from both technical and application points of view.

Nine articles have been selected based on the journal or conference associated with the article (e.g. Xu et al. (2017) published in IEEE International Conference proceedings) or based on the number of citations (e.g. Zyskind et al. (2015) has 286 citations). From these, five are about data protection using blockchain systems, and four are about blockchain architecture

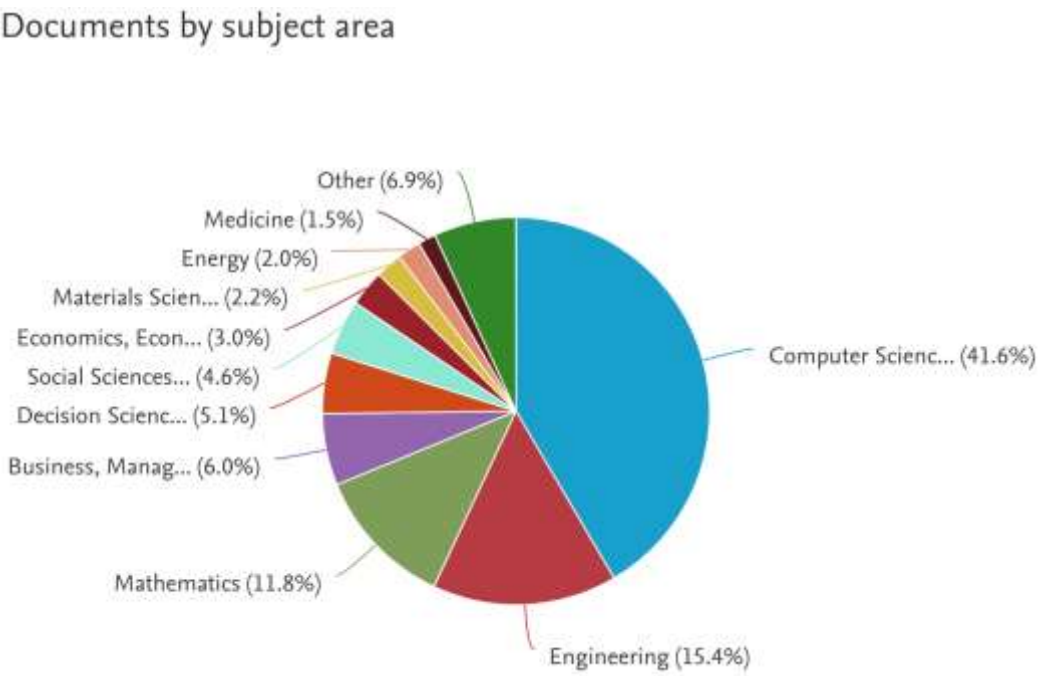

Figure 8: Documents by subject area, with the keyword "Blockchain". Retrieved from Scopus.

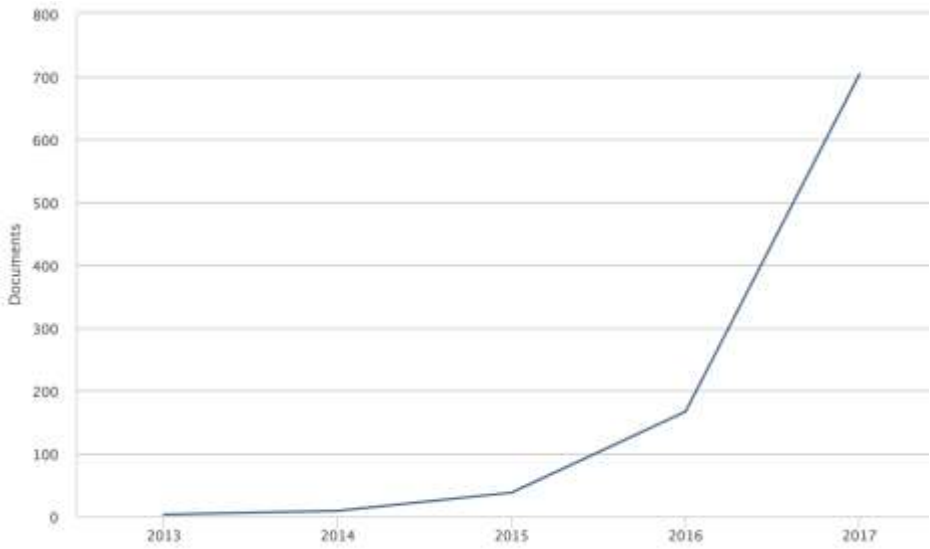

Figure 9: Scientific documents over time, retrieved from Scopus with keyword "Blockchain"

Context

Defining the context of the system is the necessary step before designing the components, according to van Engelenburg et al. (2018). It is also important to understand what the context exactly is to determine the relevance of a context-aware system for our applications. By using the keywords "context (-aware)" and "context-aware AND system" and briefly

scanning several papers, two papers have been selected for their clear explanation of what context-aware systems are and when to use these.

The paper mentioned already numerous time in this thesis (van Engelenburg et al., 2018) has been selected and describes exhaustively a method to build such a system. The paper has been suggested by the authors (of the paper) following our request to them about literature that could be useful for the design part of our research. There was therefore no online search involved for this specific topic.

Data sharing using blockchain

In this part, the solutions and information about using blockchain for data sharing are explored. We used the keywords "(Blockchain OR (distributed AND ledger) AND (data OR information) AND (exchange OR sharing OR marketplace) to retrieve 7 articles. As most articles had a recent publication date, they all had few citations. We mainly looked at the context of the publication (e.g. Karafiloski & Mishev (2017) was published for the IEEE Eurocon 2017). We also selected a whitepaper written by McConaghy (2018) as he is a recognized writer by the public blockchain community and as we met him personally at several conferences.

3.2.2 Interviews and analysis

Interviews

As in this research we are mainly involving businesses as data buyers, and as business is largely a social phenomenon, it is important to conduct interviews as part of the insights must come from practitioners (Sekaran, 2003). The knowledge acquired with the literature review is therefore complemented with four interviews. The interviewees are three stakeholders of the data exchange system and one external consultant in a research-oriented technology consulting firm. The three stakeholders are (1) one ex-privately-owned data marketplace currently transitioning towards a decentralized approach, (2) one decentralized marketplace for sensor data provider, and (3) a data supplier and decentralized data marketplace provider. The data marketplace providers that were interviewed all use different technical architectures and achieve different levels of decentralization. The table below summarizes the main information about these interviewees.

Table 1: List of Interviewees

Interviewee	Role	Country
1	Decentralized data marketplace provider	Belgium
2	Previously: privately-owned data market provider Now: decentralized data marketplace provider	Germany

3	Sensor Data Supplier and decentralized data marketplace provider	Netherlands
4	Consultant in a research-oriented technology consulting firm. Area of expertise: customer experience	Netherlands

The interviews conducted in this research are semi-structured, as there is a list of questions and areas that need to be discussed but the interviewer may ask follow-up questions or discuss another topic when he feels it is appropriate. Semi-structure interviews can provide reliable and comparable qualitative data (Cohen & Crabtree, 2006).

The complete interview protocols, including transcripts of the interview, can be found in Appendix A.

Analysis

As illustrated in the research flow diagram, the interviews provide answers that need to be analyzed in order to extract two main elements: (1) the centralized firm-related problems associated with current data exchange mechanisms, and (2) the elements which are parts of the context, and the situations that restrict the different foci. The notion of focus is introduced in Section 3.3.3 below.

3.3 A method for designing context-aware systems

3.3.1 Definition of a context-aware system

In this section, the notion of context and context-aware system are defined. The section then introduces the importance of such systems and how they can be designed using the method proposed by van Engelenburg et al. (2018). Section 3.3.2 justifies the relevance for the method for answering our research question. Section 3.3.3 offers a more extensive description of each step of the method.

In the area of information systems, the notion of *context* has been defined in various ways. According to Fischer (2012), three elements constitute the context: The stakeholders involved, the objective of the interaction, and the time and place where the interaction happens.

The term *context-awareness* was first introduced by Schilit and Theimer (1994) and referred to location and identity of people and objects, as well as the changes in these objects. It

helped them explore how a mobile application should discover and react when a change in the environment occurs. Since this introduction, the term has been vastly used in the literature on software and has been applied to more than computing applications. Baldauf (2007) describes context-aware system as increasing usability and effectiveness by giving more consideration to the context. In his words, "context-aware systems are able to adapt their operations to the current context without explicit user intervention" (p. 1).

In this thesis, the definition that will be used is a more practical one based on van Engelenburg et al. (2018). In loc. cit. a context-aware system is defined to be a system that it is able to sense and adapt, by the means of a reasoning mechanism, to the relevant part of its environment.

3.3.2 Relevance of context-awareness for decentralized data marketplaces

There are several methods to design socio-technical systems that exist in the literature, such as value-sensitive design (Friedman, 1996). For this research, as previously mentioned, we will use the context-aware design method suggested by Sélinde van Engelenburg, Marijn Janssen, and Bram Klievink (2018). In this section, we articulate the reasons behind this choice, in two steps: (1) The need for context-awareness, and (2) the choice of this specific method.

Firstly, there are many factors of the ecosystem that change values dynamically. Some of these factors and their associated values affect differently stakeholders' behavior on the marketplace. For instance, the fact that new datasets are uploaded can change the user preferences and he should therefore be notified. Context-awareness has been used extensively for building recommendation systems (Adomavicius & Tuzhilin, 2011), as also illustrated by Google search engine. Let's assume that we are focusing on the willingness of the data buyer to participate in the data marketplace. Based on interviews with an expert in IT-related customer experience (Interview 3), there is evidence that user experience on the data platform is crucial. Peer-to-peer platforms often suffer from a lack of clarity about where the information is in order to solve specific problems (Fischer, 2012). If the system understands what the context is and what the expectations of all stakeholders are, it could predict user preference and update the interface selectively, in a personalized way. For instance, if several datasets report the same phenomenon but have different quality levels, the system must present the datasets in an intuitive way for the user to know which one to buy i.e. which one has the best quality. However, as the system should be decentralized, there is nobody to check the data and rank them. The system must therefore be aware of the context and sense the data quality in order to dynamically adapt how the information is presented to the user (e.g. adapt dynamically the ranking).

Secondly, the data marketplace designed in this thesis is a structure that enables the co-creation of value by individuals, as they share information that others may not have, and in a more elaborate marketplace some stakeholders may even provide extra services such as data

curation. According to Schmidt (2002), it is crucial to have mutual awareness between stakeholders in order to support the distributed and independent activities in such value co-creation environments. The willingness to participate of data providers is also dependent on the context and more specifically their awareness of which users download their data. For instance, to improve data providers participation, the dynamic customization of sensitivity preferences should impact who accesses data. According to the data supplier interview (interview 4), based on how sensitive a dataset is for some business, access must (or must not) be restricted. Therefore, the willingness to share data by a business A depends on the sensitivity of the data and which other businesses the data is shared with. Sensitivity and businesses involved are therefore part of the context and should be taken into account when the system is asked to allow users to download this data. In addition, monetarization can influence the risk level that actors are willing to take (interview 4). If data can be sold at a higher price on the data marketplace, some actors may be willing to upload more sensitive data since the benefits outweigh the risks. This example shows that the willingness to participate is not just a yes or no decision independent of the context. There are several context elements that can influence these actors and a context-aware design method allow us to target these elements.

Furthermore, our data marketplace is fundamentally more complex than current solutions because of decentralization, e.g. actions are not coordinated, and decisions not taken, by a single party. It is even more complex within a *public* data marketplace, where there are few barriers to entry (e.g. internet infrastructure) and therefore where large numbers of stakeholders can participate. More complexity implies that there are more factors in the environment, emphasizing the need for defining which ones are relevant, as it may not be possible to monitor everything.

A higher degree of automation is also required compared with centralized data marketplaces since the platform should behave according to fair and transparent rules instead of a centralized actor's decisions. This is justified by the fact that our data marketplace aims at removing as much as possible the presence of central firms i.e. humans. However, these central actors play important coordination roles in the ecosystem, such as increasing accessibility and utility of data (van Schalkwyk et al., 2016). More automation suggests that there is less human implication in the choices related to the marketplace operations. The technology needs to "understand" the context as humans would do, but without the risks previously mentioned related to human agents owning the platform. To reach this goal of having more automation while keeping efficiency and usability, context-aware systems can be used (Baldauf, 2007). The marketplace needs to understand the relevant information in the environment and updates accordingly. In other words, the system must be aware of the context, and adapts to it. For instance, in a privately-owned platform, the owner may ask information about stakeholders who would like to participate in the marketplace and decide the access based on the information, after checking whether it would not harm public good. This is relatively easy to do in centralized settings where the firm owning the marketplace would do the checks one-by-one. In a decentralized data marketplace, the entry verification to the marketplace has to be done automatically or by third-parties, as there is no controller of

the platform. Fundamentally, the central governance previously operated by a firm is now operated by the technology (or by the third-party, but this is also some form of firm in a central position, with interests that may differ from public good).

Finally, one of the properties of blockchain technologies is their immutability, or the fact that the content which is written should in principle not be possible to change. We mentioned "in principle" because although some events in the past have led to some blockchains developers to remove transactions using a "hardfork", this practice is highly cumbersome and should be avoided (Lin et al., 2017; Wong, 2016). As transactions cannot be reversed, this increases the damages in case of a wrong manipulation by the system. It is therefore necessary to minimize these wrong manipulations. To reach this objective, it is also necessary to have a sufficient understanding of the context in order to be able to sense events and react to stimuli according to rules that have been pre-defined.

Now that we have justified the need for a context-aware marketplace, we provide explanations for the choice of the design method developed by van Engelenburg et al. (2018).

The first reason is that this method provides the reader with an exhaustive plan to build the system, including a detailed description of each step. Following the method is therefore sufficient and no other research about designing context-aware systems is necessary. In addition, the method is especially targeting designers. As a consequence, the paper does not go too deep in presenting logic theory, neither does it lose the designer by defining terms in a very abstract manner. It actually provides very clear definitions including concrete criterions, and support these by giving numerous application examples. The step-by-step description and clear designer targeting greatly help the designer to quickly understand the concepts and method instead of confusing him. As a result, the designer can fully focus on the application of the method.

Finally, a crucial point to justify the use of this method is the support from the authors of the method for this thesis. The authors were able to provide advices about how to use the method, and feedbacks about the application of this method to the data marketplace. It should also be noticed that the authors have already applied the method to a blockchain-based information sharing system between businesses and governments, in the case of customs.

All these reasons have led us to select and apply the context-aware system design method developed by van Engelenburg et al. (2018). This method if described in Section 6.2, and applied from Section 6.3, until the end of Chapter 6. The method, its application, as well as the results, are then discussed in the discussion section in Chapter 8.

3.3.3 Basic requirements or context-awareness

Concerning the design phase, not all parts of the system need to be context-aware. There are some universal facts that need to be considered in the basic requirements, however they are

not context-aware. As an example, saying that the marketplace should demonstrate a "good customer experience" is not related to context-awareness. There is no case where the designer wants any of the stakeholders to have a bad customer experience, and there is therefore nothing in the context that should influence (dynamically) the customer experience proposed by the marketplace. Therefore, in Section 6.2 we first suggest the basic architecture that the system should have, based on the requirements, the technologies available and the designer goals. Following this, the method is described and then applied.

Chapter 4: Data sharing

4.1 Introduction

In this section, insights from the literature review and interviews are explored. Section 4.2 will introduce data as a key element for business intelligence and for machine learning. It highlights the need for sharing data. Section 4.3 then describes the three main ways of sharing data: open data marketplaces, data brokers, and privately-owned data marketplaces. In section 4.4, the centralization-related (i.e. firm-controlled) problems are articulated by identifying the different parts of the data exchange process that suffer. Section 4.4 describes blockchain technology.

4.2 Data importance

Data has become a valuable asset for our modern society, leading to innovation and economic growth (Zyskind, 2015) for many organizations and across all kinds of industries. Decision-makers use datasets for more efficient governance solutions (van Schalkwyk et al., 2016), researchers find correlations and make science progress, businesses improve their forecast thanks to business intelligence and analytics. Advances in computer technologies such as network connectivity and disk storage, which have therefore become more affordable, have played a significant role in the production of data (Sweeney, 2002).

Firms have realized that data could bring a competitive advantage and have thus brought their data capabilities forward, storing more data than they actually use, hoping that they will eventually make use of these (Interview 3). Some firms even have data as the core of their business model, such as Google or Facebook that have both reached significant user bases and market capitalization (Haucap & Heimeshoff, 2014). This has led to the definition of new terms such as "Big Data", a reference to massive datasets that require sophisticated tools for data manipulation (storage, management, analysis, and visualization) (Chen, 2012), or "Business Intelligence & Analytics (BI&A)", a data-driven decision-making approach for businesses. To understand the potential of our sensor data for businesses, it helps to understand the impact of their analysis through BI&A.

Business Intelligence & Analytics

BI&A was popularized in the 1990s and grew in three waves, each characterized by new enabling technologies and a shift in data collection paradigms. The BI&A 1.0 was the result of improvements in data warehousing and relational database management systems containing mainly structured data (Chen, 2012) i.e. data with a given format. Then the internet offered new opportunities in terms of data collection, with the rise of some of the huge "dotcom" companies. This BI&A 2.0 wave was in particular focusing on retrieving

information from web content which is often presented in unstructured ways. Finally, the BI&A 3.0 encompassed "person-centered, context-relevant data crawled from devices" (Chen, 2012). Autonomous applications such as mechanical robots and chatbots can interact with users to create data. New technologies for data collection have enabled each of these waves and it remains to be seen which ones the catalyst for BI&A 4.0 could be.

The impact of business intelligence is widely spread across all industries and sectors, e.g. science & technology, healthcare, and politics (Schneier, 2015) which require different levels of security depending on the sensitivity of the data. As an example, healthcare usually deals with sensitive data and there are therefore strong regulations (e.g. HIPAA, IRB) that parties have to comply with to ensure privacy and ethical research (Gelfand, 2012). In part because of the stronger regulations and more sensitive data, health data analytics face more difficulties to scale and is therefore less developed than other commercial applications (Miller, 2012). This example illustrates that industries where access to data is hindered demonstrate less innovation and emphasizes the importance of data.

Machine learning and democratization of AI

There is also much demand for data in the field of machine learning, where algorithms are fed with very large datasets to forecast future behavior when presented with unknown data. Halevy et al. (2009) has written that "Simple models and a lot of data trump more elaborate models based on less data". This confirmed the findings of Microsoft's researchers Banko and Brill (2001) illustrated on Figure 11 below. We can see that more advanced models improve the test accuracy by some percentage (depending on the inputs), but the impact of feeding models with very large amounts of data is much stronger, with increases of more than 20% in some cases when increasing the number of data points from less than a million to a billion.

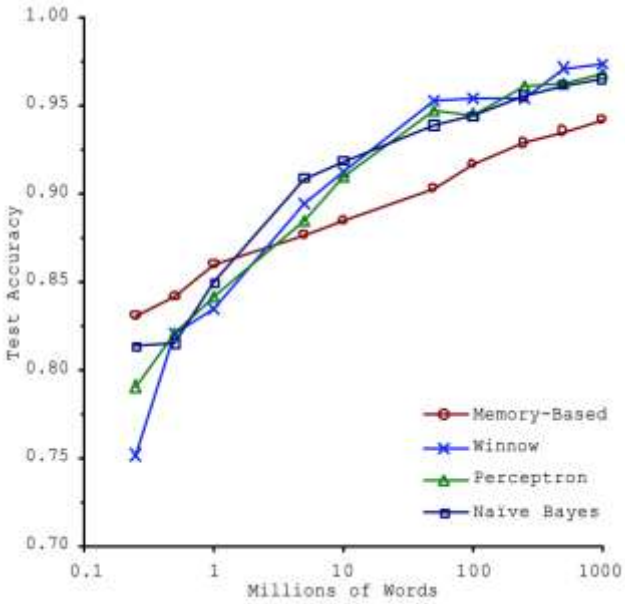

Figure 10: Task performance for different models and increasing size of datasets (Banko and Brill, 2001)

Machine learning is an area that is developing fast and the applications are numerous. However, not everybody has access to data for training their models. The field is led by a few companies that have both advanced models and especially large datasets. These leading companies such as Google or Facebook often offer services that collect massive amounts of data about users (often in exchange for free services) and store these in data centers. Smaller startups face difficulties to access this data and therefore face high barriers to become competitive in the AI market (Interview 1). Roman & Stefano (2016) also mention that there is a lack of data in the small and medium enterprises segment, resulting in the need for making approximations.

In parallel, there is increasing awareness about the need for democratizing AI (as illustrated by the Association for Computing Machinery's AI Decentralized Global Initiative). This does not mean democratizing the access to AI by offering cloud services, but rather democratize the possibility to develop AI algorithms by enabling access to data. Some researchers even go further and strongly recommend to not have AI development in the hands of a few players in order to reduce potential existential risks (Bostrom, 2003), but discussing these theories is out of the scope of this research. The conclusion of this section is that there is scientific support for the (potential) impact of data and the need for increasing its access and exploitation.

4.3 Current data exchange mechanisms

The introduction of the thesis briefly introduced the *silo problem*, which described the fact that despite the need for data and its increasing supply, most data remains unused and is just stored (Interview 3). For instance, according to a study from McKinsey & co. (Deichmann et

al., 2016), one oil rig with 30,000 sensors uses only 1% of the data produced to detect anomalies, "ignoring its greatest value, which involves supporting optimization and prediction activities." The literature discusses how to break this data storage problem and reusing it across organizations for commercial purposes.

The solutions offered in terms of data monetization are still limited, despite data being increasingly characterized as an asset (Schwab, 2011). Multiple industries require data exchanges and solutions are being proposed to accommodate this need; studies have shown that more than half of the individuals would sell some of their data (Parra-Arnau, 2017). Those findings combined with the data markets survey (Stahl, 2014) indicate that over the last few years data commercialization has gained attraction. Several factors can be assumed to be responsible for the shift in paradigm towards more data sharing and open data: increasing awareness of data value; surging amounts of data produced by what has become a "digital society" (Interview 3); and the evolution of IT infrastructures to collect, share, and store data. In addition, the opening of firms' boundaries for innovation, which has led stakeholders to co-create value using both internal and external resources (Chesbrough, 2006), may also have had a significant impact.

Before discussing the different possibilities to commercialize data, we first give an overview of the stakeholders interacting within the sensor data system, as described by the literature. The data system is made of data providers, users, custodians, managers, and service providers. Data service can further be decomposed into different kinds of data-based services. Such services include for instance: sales (data brokers) or data gathering (data aggregators). After introducing the stakeholders and the related vocabulary, we discuss the silo problem. We then elaborate on how the actors interact to exchange data across three main data exchange mechanisms that have been proposed: data brokers, privately-owned data marketplaces, and open data marketplaces.

4.3.1 Data exchange ecosystem description

Data providers

Data providers supply data, generally in a raw format, that they collect as their main objective (e.g. meteorological measurements) or this data is generated while operating their main activities. As an example of the latter case, telecom companies provide communication services as their main function, but also generate data such as the location of mobile phones, which can be used for other applications such as public security.

Usually, data providers have an economic incentive to provide data (Gopalkrishnan et al., 2013). The benefits are higher if the related industry has fragmented information, if sharing information is not a risk for the business model, and if the data is unique and scarce (Banerjee et al., 2011). With increasing competition, data providers have to differentiate themselves (Koutroumpis & Leiponen, 2013), by for example adding extra services upon their data directly (e.g. curation service or insights).

Data users

Data users need data for various purposes such as decision-making, business analytics, or training predictive models, as explained in section 4.1. An additional concrete example of a sensor data usage is an energy company that may be looking for sensor data owned by households in order to get insights into the electricity consumption in real-time. These insights could be used to increase shortage predictability and provide a more robust electricity grid (Mengelkamp et al., 2018). Several other examples have been given in the introduction of the thesis, about how self-driving car developers crucially need very large amounts of data.

Data managers

Leiponen et al. (2016) describe data managers as organizations that catalogue and clean raw data to improve its interpretation. Therefore, they add extra value to the raw data by improving efficiency in the ecosystem (Zhu et al., 2010). Van Bommel et al. (2005) provide some examples of services that can be provide such as data curation (formatting, identification of outliers, and language translation).

Data custodians

Data custodians are stakeholders that provide a 'trust' infrastructure (e.g. cloud infrastructure) upon which the data is stored and can be accessed (e.g. via APIs). However, these infrastructures may not be that trustful according to some sources (Roman & Stephano, 2016; Interview 1, 2, 4). This will be further investigated in the next section as the relation between trusted infrastructure and intermediaries. Data custodians may also offer complementary services such as auditing and certification to ensure data quality (Perrin et al., 2013) or products which give the possibility for data owners to control their information (Eggers et al., 2013).

Data aggregators

Data aggregators are specific service providers who collect and aggregate data based on one specific sector or application. As an example, Amadeus gathers data provided by airline companies to provide websites such as skyscanner.net with the access to the aggregated data. These websites access APIs and present the results in a user-friendly way to customers as demonstrated by providing a comparison or recommendation service. As a result, customers can instantaneously see which deals the best are without having to check each air flight company websites individually.

4.3.2 Data sharing for small data exchange ecosystems

Pasquetto et al. (2017) defines data sharing as "the fact of releasing data in a format that can be used by other individuals." Based on this definition there are a multitude of ways to access data. Data can be shared via private exchange between researchers, such as posting datasets on researchers' websites, depositing datasets in archives, or attaching data with publications.

Although these exchange mechanisms may be useful in very specific research areas, it is difficult to imagine how such systems could scale as both parties need to know each other to some extent and find data by checking several research portals. It is therefore a mechanism that could work in research, but other solutions need to be implemented for large-scale sensor data sharing with stakeholders other than researchers and their related visibility (thanks to their publication). In the same way, sensor data providers and businesses could also exchange data using traditional tools such as listing API access or datasets on their website or using common communication tools such as email to send datasets. Again, this research focuses mainly on the possibility to have data exchanges from many to many while minimizing the search costs.

Three solutions presented in the literature have been selected as potential ways that can scale by being efficient ways of sharing data: data brokers, privately-owned data marketplaces, and open data marketplaces. In section 4.3.3 to 4.3.5, we review the literature to answer the first sub-question with the additional scope adjustment about scalability: *What are current scalable solutions used for data sharing between sensors owners and data users?*

4.3.3 Privately-owned data marketplaces

Marketplaces are places where suppliers and customers can meet each other (Henderson & Quandt, 1980) to indicate their intention to buy or sell certain products which eventually match and may be settled (Schmid & Lindemann, 1998). They reduce search costs by efficiently listing data and can increase revenues by using product recommendations (Subramanian, 2018). The term *privately-owned* refers to the fact that the data marketplace is developed and owned by a company. Amazon and eBay are examples of e-marketplaces, mostly for physical items but increasingly about digital assets as well (e.g. Amazon is providing e-books). In a data marketplace, providers supply the marketplace with two types of data (Interview 2): (1) data streams, via an access to the API of the sensor owner in (almost) real-time, and (2) datasets such as historical values of a sensor (e.g. acceleration of a self-driving car based on the position of the closest car). Usually, data suppliers propose their licensing conditions or follow a benchmark given by the platform provider, in addition to the global terms & conditions defined by the platform (Deichmann et al., 2016). Stahl (2014) indicates the common data formats: XML, CSV, and JSON. Stahl (2014) also studied the trustworthiness of data vendors on marketplaces, according to data users. While 80% are classified as moderately to highly trustworthy, the remaining 20% are considered as showing low trustworthiness. These findings are important when considering quality issues on the marketplace, since if we remove the central firm that can check quality, quality must be

assumed based on what we know about data suppliers and is therefore about trust. Our system will need some mechanisms to support different formats (interoperability), incentivizing providers and users to adapt to these formats, and ensuring some quality check to avoid problems related to low quality of the bottom 20%.

Deichmann et al. (2016) complements these insights by suggesting four potential roles of the data marketplace, based on its maturity (its user base). If the marketplace is in an early phase, it will act mostly as an intermediary to exchange raw data and to some extent defines standards (e.g. format). In more mature versions of data marketplaces, more value is added by providing a data-as-a-service ecosystem. For instance, by aggregating data (e.g. by region) and verify the quality. In other words, within a marketplace, data can be bought by some parties that can then post value-added solutions in addition to the raw data offered in the marketplace.

Concerning the structure of data marketplaces, Deichmann et al. (2016) have proposed six guidelines for Internet-of-Things data marketplaces: (1) building an ecosystem, (2) opening up new monetization opportunities, (3) enabling crowdsourcing (i.e. giving incentives and keeping the access open for more data suppliers to participate), (4) supporting interoperability (cross-device, cross industry), (5) creating a central point of "discoverability, and (6) achieving consistent data quality. These key elements are important and will be used in the design phase. In addition, they specify that "the data marketplace needs to assume a neutral position regarding participants."

Researchers have studied valuation methods for selling proprietary data in marketplaces (Parra-Arnau, 2017). There are typically two monetization models: subscription service (i.e. pay per volume or duration) and on-demand (i.e. pay per use). An alternative model is the "give and take", where customers can access data only if they have provided useful data. However, no instance of this monetization model has been found as an available product and little literature has been found to explore this opportunity.

There are several examples of data marketplaces already available, such as Infochimps (acquired by CSC) or Microsoft Azure Marketplace.

4.3.4 Open data marketplaces

Several definitions of *open data* can be found in literature. Janssen et al. (2012) defined open data as "non-privacy restricted and non-confidential data that is produced with public money and is made available without any restrictions on its usage and distribution." According to the definition given by the OpenKowledgeFoundation (2005), the absence of profit generated from it is a core principle of open data. Some sensor data could be classified as open data e.g. weather properties are not secret and can be measured by anyone equipped with the right sensor, there is therefore no privacy nor confidentiality. If these data are also freely available

for any use, they qualify as open data. This section will therefore investigate the literature about open data marketplaces and extract information for sensor data marketplaces.

In the past years, new directives on public sector information sharing in the EU and the open government initiative in the US, as well as public demand, have led to the public sector sharing more data. Open data is considered as contributing to digital service innovation and therefore to economic growth (European Commission, 2011). It empowers individuals, both data users and providers, to participate in the data economy (Neuroni et al., 2013) and therefore creates job opportunities.

As defined above, much data used by businesses would not be qualified as open. In fact, companies may also use privacy-restricted and confidential information which could be made available with restrictions on usage and distribution. The absence of restrictions on usage and distribution makes the theoretical monetary value (on a market) of these pieces of information worth zero and business models cannot emerge based on selling that data (unless funded by public money). Nevertheless, getting insights into open data marketplaces helps us understand the broader data marketplace ecosystem and in particular how a marketplace impacts the barriers to data sharing.

In their paper "Exploring the Value Proposition of an Open Data Marketplace", Smith et al. (2016) use a case study approach to define the elements of data marketplaces and evaluate its contribution to the open data ecosystem. Open data marketplaces are defined as "intermediary platforms which provide the requisite infrastructure, rules and services for transactions of open data, knowledge and experiences between open data providers and users" (Schmid & Lindemann, 1998). The main role of such a platform is therefore to simplify the process of exchanging open data and data services in an open cooperative environment (Zuiderwijk et al., 2014). The main elements of the marketplace studied by this research are the following and complement the findings from section 4.3.3. These elements will be considered again in chapter 5 in the design phase.

1. A technical platform with the back-end, an API management system, and statistics gathering. The API management systems grant access to the API, with some conditions such as a limited access to avoid the system being overloaded with unnecessary requests.
2. A website with AP descriptions, news feed, data catalogue, and other information which can facilitate the user-experience.
3. A support service which checks data quality, as well as provides documentation and templates; operational status; and help.
4. Knowledge sharing activities such as meetups and project showcases.

Digital marketplaces result in five main values (Smith et al., 2016). First, it lowers task complexity thanks to the centralized website; users do not need to browse independently to each data provider's website to collect data. They can also easily compare the products (the APIs). Second is the higher access to knowledge. In addition, there are increased possibilities

to influence, as the data brokers which manage the data marketplace receive specific requests from data users and transfer to data providers, fostering innovation. Fourth, it lowers risk because users can see that the APIs are used by others. This "community aspect" increases confidence in the dataset. Finally, there is a higher visibility of the data. It is also important to note that marketplaces are subject to network effects, mainly positive (Eisenmann et al., 2006), meaning that the marketplace value increases for both sides when the number of suppliers and users grow.

Impact of open data marketplace on open data adoption

There are six main barriers to open data adoption (Smith et al., 2016):

1. Institutional barriers (e.g. accountability versus entrepreneurial spirit by firms)
2. Task complexity barriers (e.g. finding and analyzing data)
3. Use and participation barriers (e.g. lack of incentives and time)
4. Legislation barriers (e.g. privacy, disputes resolution, contracts)
5. Data quality barriers
6. Technical barriers (e.g. lack of standards)

The findings of the case study have resulted in positive impact of open data marketplaces on open data adoption for 2 to 6 and no impact on the institutional barriers. Interestingly, they have also demonstrated that there are also negative impacts on use and participation barriers. According to Smith et al. (2016), the former is explained by the fact that users perceive the centralized API management system as not scalable nor sustainable in the long-term. They claim that there are risks of overloading the platform because the provider is proceeding on a request-based API instead of allowing bulk downloads. As a consequence, data users are reluctant to build data service innovation based on the system as provided i.e. they are not willing to engage in building business models around data-related activities if they do not believe that the platform can reach a large customer base (where customers are both data providers and users). We conclude from this information that scalability of the marketplace is crucial to motivate data service provider to participate in the marketplace. An instance of data service provider building a business model around it are consulting companies writing data-driven reports for their clients.

Finally, in such a system the platform manager decides upon who can access the data and may ask concrete information such as the number of end-users of the data user or more details about the projects who want to use the data. It is difficult to know whether the managers are biased towards some data providers or not, which is not a huge deal in an open data ecosystem but would become when financial, private, or competitive components are added.

As a conclusion, on the one hand open data marketplaces bring mainly positive contributions to the open data ecosystem by lowering barriers to adoption. On the other hand, they also have some negative impacts on use and participation as well as data quality barriers because of the increasing perceived distance between data users and providers and because of the doubts about scalability of a centralized API management system.

4.3.5 Data brokers

Another typical structure for exchanging data is using a third-party or "data broker" without marketplace. Data brokers collect data from data suppliers and resell or share with third-parties. They act as an enabler for exchanging data to some extent as they can connect the supply and demand sides. In addition, they can also sell to other data brokers and thus merge different networks. They may also provide extra benefits such as verifying identities of stakeholders in the network (both from the supply and demand sides) or act at the same time as data managers and aggregators. In addition, as they may have access to various sources of the same datasets, they can corroborate the data to increase their validity. As a consequence, it can help prevent fraud, improve accuracy of data in circulation, improve product offerings, and tailor advertisements (US Federal Trade Commission, 2014).

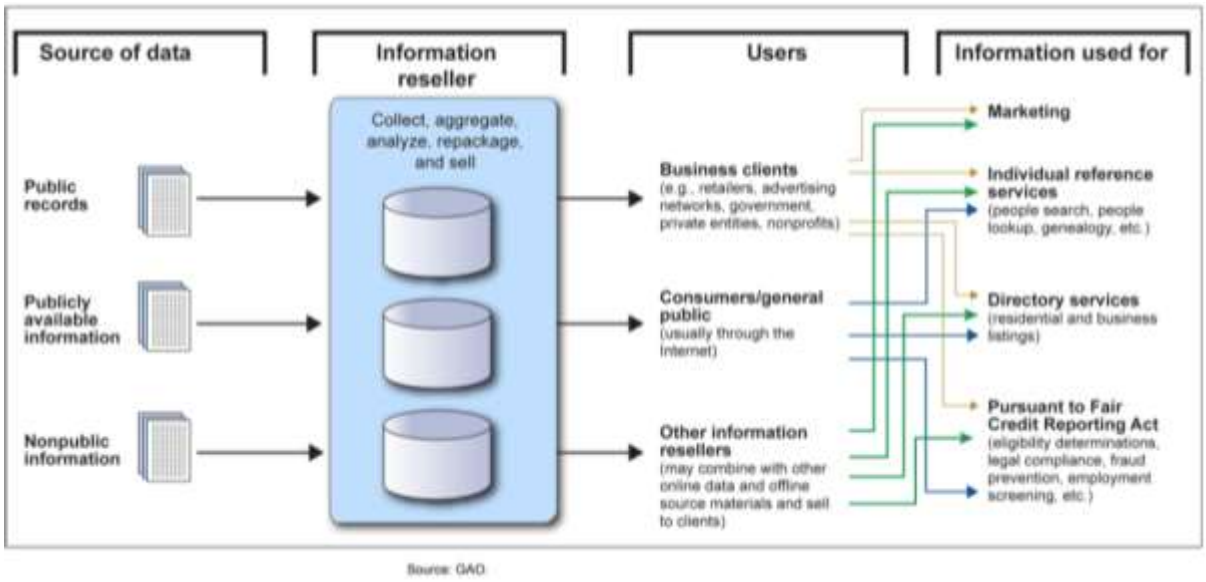

Figure 11: Typical flow of data from sources to data users, through data brokers (US Fed. Trade Comm., 2014)

Figure 12 represents the flow of data through data brokers. Data brokers maintain large organized databases where they store the information they collect, from customers directly or via other data brokers. The information is collected in the first place from three primary types of sources (US Government Accountability Office, 2013): public records, publicly available information, and nonpublic information.

1. Public records are documents generally published by governments.

2. Publicly available information is not obtained from governments but is still publicly available via public services such as phone repositories or newspapers.

3. Nonpublic information belongs to "proprietary sources" (US Government Accountability Office, 2013). This may include information from businesses such as via social networks or surveys.

In our case, sensors data are mainly part of the proprietary source. Nevertheless, it is also possible that some public services collect data via sensors and make it available not only on open data marketplaces but also via data brokers.

The data broker model has raised concerns in terms of transparency and accountability (Federal Trade Commission, 2014) and this study about data brokers has led to the following observations:

There is a lack of transparency about which data is accessed

Data brokers gather data from a large number of sources without customers' consent and even without them knowing about it. This can result in a consumer being refused a transaction based on a wrong assessment by the data broker, and the customer may not find the source of the decision. Data brokers may also proceed to sensitive inferences based on the data they have. For instance, here the sensors in the car could be connected to the car company that resells data to insurance companies without users' consent, leading to increases in insurance plans if the car owners' driving behaviors are judged dangerous.

There is a lack of transparency about who has which data

The data broker industry is a complex network of actors providing data to each other. Once the information enters the data brokers network, it can be quickly diffused with others. As a consequence, there is a lack of information about this data reseller industry, making it difficult to estimate its size and which types of data are shared.

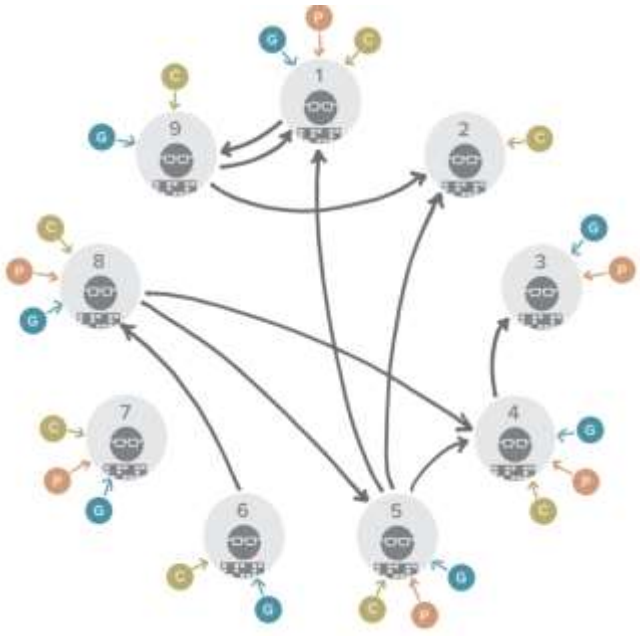

Figure 12: Data brokers form complex networks of information diffusion (US Fed. Trade Comm., 2014)

Data may be stored permanently, leading to more risk exposure. The storage tools are not optimal.

Some data brokers store data indefinitely about users, even in cases where it is not necessary. This represent a risk as the data is exposed to potential attacks for a long time. As technology evolves, effective cybersecurity protection strategies today may not be that effective in a few years. As an example, breakthroughs in quantum computing could break traditional cryptography techniques and therefore make potentially all encrypted data readable (Bernstein, 2009). Intermediaries currently host data from providers, potentially unencrypted (Interview 1, 2, 4). Generally, these are stored on centralized servers or clouds. For the servers, this implies that the system has single points of failure which if exploited may lead to massive data breaches. As a consequence, researchers have recently suggested distributed databases systems (Bulkowski et al., 2018) and we will discuss further this use in the design part. Data can also be stored and replicated in the cloud to improve protection against involuntary data deletion. Nevertheless, such solutions still imply significant power of a single actor (i.e. the cloud provider). In practice, many firms do not trust cloud providers and would not upload their most sensitive data on the cloud (Kshetri, 2013; Qian, 2018).

Guidelines have been proposed (Federal Trade Commission, 2014) in order to limit the impact in the case of some data brokers or other stakeholders behave maliciously within this ecosystem. These guidelines include the implementation of privacy-by-design, or the consideration of privacy at every step of the product development. In order to protect privacy, some data brokers use anonymization techniques such as K-anonymization (Sweeney, 2002) which is a widely used privacy-enabling technique that allows sharing databases while keeping some degree of privacy about each individual item belonging to this database, all

while the data remaining useful. Nevertheless, researchers have proven that it is possible to de-anonymize the datasets (Narayanan, 2006).

Despite the efforts in the industries to make it more transparent through extensive regulation and self-regulation implementation, the Federal Trade Commission of the United States claimed that there is still a lack of transparency about data broker practices (US Government Accountability Office, 2013) and data leakages led to new investigations about their practices. Public organizations such as the US Congress or the EU parliament impose more restrictions to maintain privacy, which reflects the evolution in methods of data collection. The dilemma is how to set up these constraints while avoiding inhibiting commerce and innovation (GAO, 2013

4.4 Relying on trust in a centralized exchange: a barrier against data sharing

In the previous chapters, we have observed that the data broker model involves a third-party getting the data from the provider and selling it to the users. This data broker stores data in data storage system that he owns and, as indicated on the diagram below on Figure 14, mediates the data transfers.

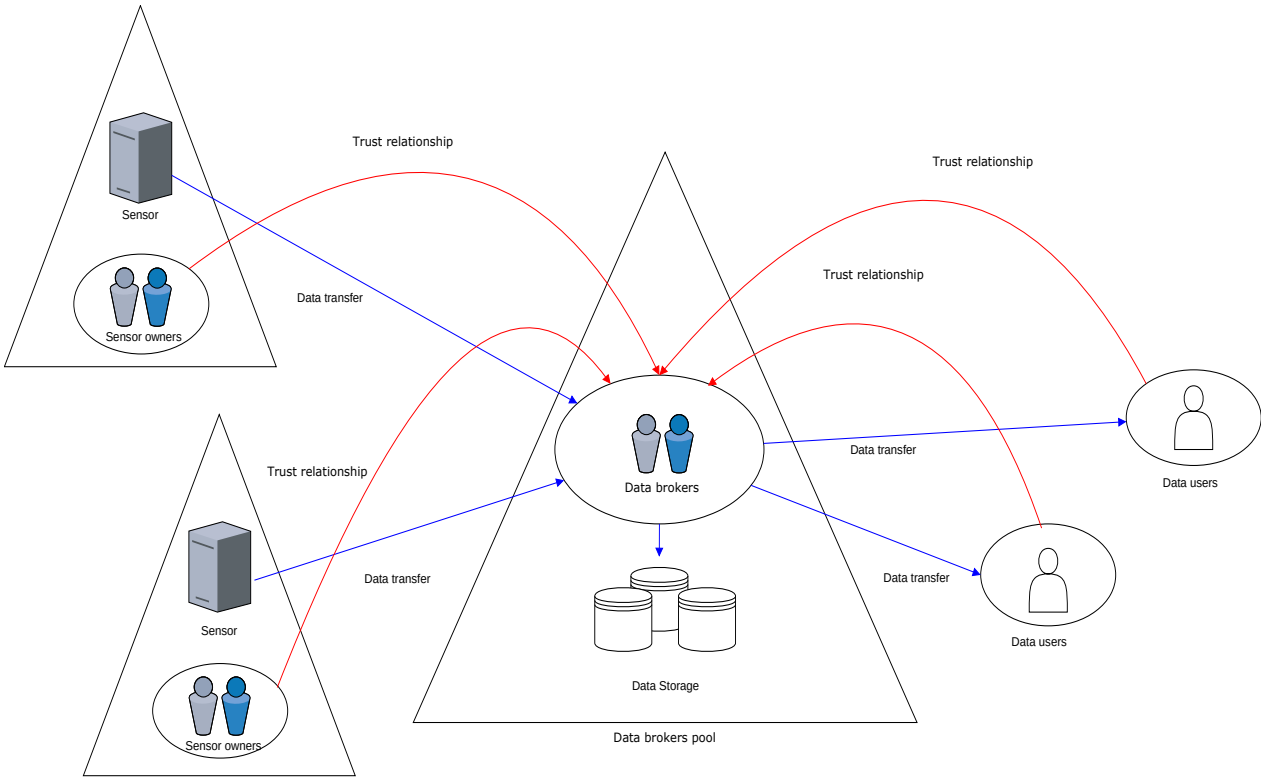

Figure 13: Stakeholders have to trust data brokers in data exchanges

This system involves a great amount of trust in data brokers or firm controlling data marketplaces, from both sides, represented by the red arrows. Sensor owners trust the centralized company for not sharing information with the wrong stakeholders (e.g. a car driver may not want his data to come into insurers' hands) and data users want data brokers to actually provide them with data and not run away after the payment, neither do they want to be provided with low quality data. The trust problem is a broad term: it denounces various types of trust drawbacks, such as technical ones. For example, a trust-related technical problem would be that the data is stored on the database of one actor who can decide on their usage, from sharing to deleting.

As also mentioned, the lack of transparency of data brokers has raised doubts about risks related to their omnipresence in data exchanges. Jahansoozi's work (2006) on organization-stakeholder relationships reminds us that transparency is a key element for establishing trust within these relationships. Roman & Stefano (2016) confirm this statement: "All of the data marketplaces are essentially centralized systems, where participants in the marketplace have to trust a third-party with managing their data." As a consequence, the company behind the data marketplace could share the data with other third-parties, use the data for its own benefits, or even delete data that is not used "often enough" to save some storing space, in the cases where they store data. Storing data is in their interest since storage space is an extra revenue for the company and since it is more cost-effective for other firms. Typically, the data is "protected" by terms & conditions and privacy policies that are often vague, resulting in data owners losing control over their data. (Roman & Stefano, 2016).

The lack of transparency is not only about data protection. There is also no guarantee that the company is not pricing the data dynamically to maximize its profits, tailoring the price to customers (Subramanian, 2017). In addition, the pricing reflects the transaction costs, which are higher if there are several companies involved as intermediaries or these companies have more power in the system, as explained by Porter's Five Forces (Porter et al., 1985). For instance, Figure 15 represents a situation in which a car owner produces some data that are then sold to a data user after going through a cascade of data brokers that each take a fee, resulting in 40% of the payment only going to the data provider.

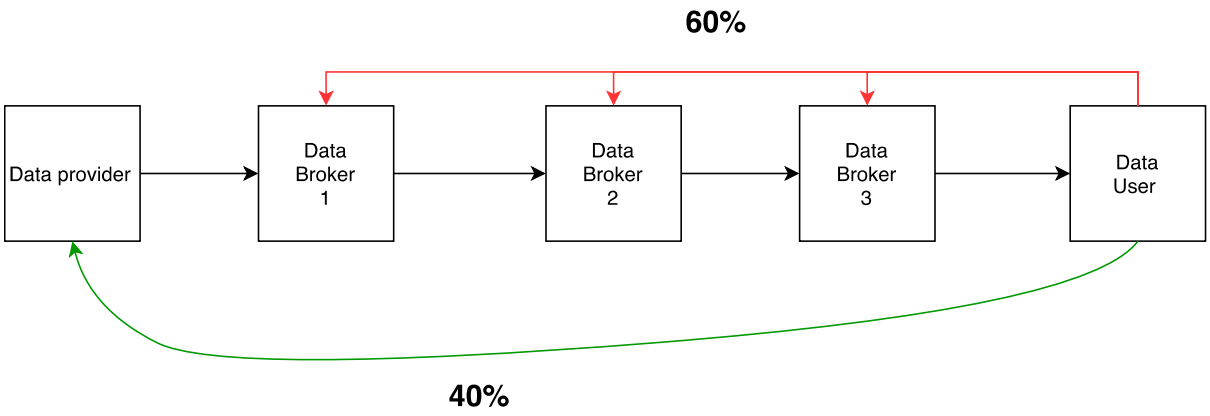

Figure 14: Data chain, from provider to user.

Hoffmann et al. (1999) have stated the famous quote "Trust is a crucial element in any kind of transaction between a buyer and a seller on marketplaces platform." Third parties have therefore been introduced to mediate transactions between buyers and sellers as a way to 'ensure' trust in the transactions. They were assumed to behave ethically and provide users with strong data protection. Nevertheless, several cases have shown that these were not always respected, as illustrated by the recent Facebook data scandals e.g. the Cambridge Analytica case (Cadwalladr et al., 2018). Facebook is a form of data broker as it collects data about users and provides data-services to organizations. The following quote found on the Ocean Protocol Foundation website summarizes salient critics of centralized exchanges.

"Centralized data exchanges fail because they lack fair and flexible pricing mechanisms, data providers lose control over their assets, and there is a lack of transparency in how the data is used. So, data remains locked up due to a lack of trust." – Ocean Protocol Foundation. Retrieved from https://oceanprotocol.com/

This lack of trust in centralized data exchanges is among the root causes for businesses to not share their data (Interview 1; Roman & Stefano 2016). Individuals tend to share more data when they trust the parties they are sharing the data with. This implies not misusing data including sharing with others, nor finding themselves facing opportunistic behaviors by the party they have shared the data with. (Hart and Saunders, 1997). When sharing information, stakeholders generally lack knowledge about other participants and their reliability (Mishra, 1996), which is emphasized in complex networks like the data exchange industry which could be operated on a global scale. As a conclusion, there is a need for an architecture which enables the creation of a data marketplace without having to rely too much on trust in other individuals. Not relying on trust encompasses that there should be limited third-parties interacting in the exchanges of data between supply and demand.

As a consequence of these questions about the place of a centralized firm with commercial incentives within exchange systems, ways to replace these have been explored (Subramanian, 2017; Ocean Protocol Foundation, 2017; DatabrokerDAO, 2017). Specifically, despite having different approaches, several projects have a characteristic in common: they all explore the use of *distributed ledger technologies* for building their data marketplaces., The following section discusses extensively blockchain technology, an instance of distributed ledger technologies, and how it is used to reduce the central point of control.

Finally, an important note is that not all firms acting as third parties are irrelevant and should be removed. In the case above presented in figure 15, the data brokers could be (1) a gateway provider sending the data from the sensor to the car company (usually telecom companies), (2) the car company checking the results and assessing whether the data are accurate, or (3) a data aggregator that makes a report about driving-related points that users may be interested in. In this example, all third parties bring some value and it is therefore normal that the data user pays more than the price of the raw data to compensate for the extra services. Therefore, chains of intermediaries will hopefully still exist, but it is necessary to keep track of the data, for the data provider to know who is using the data and for which purpose, and for the data user to know where the data come from. Transparency is one of the main reasons required from both sides in a marketplace, according to customer experience experts and data marketplace providers (Interviews 1, 2, 3, 4).

Chapter 5: Blockchain technology

"A set of emerging technologies in the context of cryptography, cloud, and decentralized computing, such as Blockchain, Smart Contracts, homomorphic encryption and multi-party computation, offer a unique opportunity for the creation of a trusted ecosystem where large-scale data sharing can be enabled". – Roman & Stefano (2016, p. 99)

5.1 Introduction

Distributed ledger technologies have emerged in the past years, especially under the name of blockchain technology which first notorious implementation was the Bitcoin (Nakamoto, 2008). The main high-level purpose of blockchains as initially invented are their ability to solve the double-spending problem, thanks to a cryptographic proof instead of trust in payment gateway. Blockchains are essentially a mix of existing knowledge such as cryptography and distributed systems which when brought together offer systems a set of properties like immutability, transparency or absence of a single point of failure. Don Tapscott & Alex Tapscott (2016), authors of the best-seller book "Blockchain Revolution", claim that blockchain will bring a complete change in society by disruptive many industries, by for instance offering access to banking for the unbanked or enabling decentralized autonomous organizations using tokenization. The blockchain main characteristic is that it is a peer-to-peer network, without central authority to control the blockchain. As described in the research approach chapter, blockchain technology has therefore a high relevance in our system to remove the centralization of control.

Some of the typical values that blockchain technologies allow are decentralization, transparency, and immutability; and therefore censorship-resistance, auditability, and efficiency across extended geographic locations.

Section 4.2 describes blockchain on the technical level, before articulating how it can perform the aforementioned values. Section 4.3 then gives an overview of blockchain importance in (data) marketplaces, and section 4.4 introduces some blockchain-based mechanisms that will be needed in the design phase such as the token curated data.

5.2 Technical overview

A blockchain is a *data structure* composed of blocks that contain information and are linked together, such that a change in any of the blocks can be noticed rapidly. This whole data structure acts as a distributed ledger, enabling the transfer of assets without the need for intermediaries as the technology prevents *double-spending* (Nakamoto, 2008). Double-spending refers to an old problem in digital payments which is the possibility of duplicating

assets and to spend them twice, by taking advantage of a lack of synchronization of stakeholders within the network. However, Blockchain allow the stakeholders to agree upon the state of the network by providing the distributed system with a *consensus algorithm*.

The first consensus algorithm proposed within blockchain applications is called *proof-of-work* and consists of some of the participants of the network giving up hashing power (i.e. electricity consumed by proceeding to mathematical computations). Several consensus algorithms have been proposed to cope with the scalability issue of proof-of-work (the fact that it consumes so much electricity makes it unusable on a large scale). The blockchain trilemma (Buterin, 2016) says that no consensus algorithm (in the context of blockchain technology) can achieve more than two of the three values: decentralization, scalability, and security. However, these properties are not binary values, meaning that we could reach a combination of levels of these that is satisfying for the expected applications. Discussing this trilemma is not only out of the scope of this research, but also a major research topic with considerable resources invested into it. Nevertheless, we introduced this trilemma because the level of decentralization that we would like to achieve will be discussed. Whether we can completely remove intermediaries in the system or not will be judged based on the contextual information that we collected and that were described in the previous section.

The blocks include transactions (e.g. payments, transfer of ownership, land registering, or something similar depending on the application), which are connected to each other via pointers ("hash functions"). The particularity of these hashes is that if the input is changed even a tiny bit, the resulting output of the hash function is completely modified, resulting in a directly detectable modification by the network and such change is refused by the network keepers if appropriate.

There are four main steps: (1) generating a pair of private-public keys and creating a transaction, (2) broadcasting the transaction to the network, (3) adding transactions to a block, and (4) linking the new block to the blockchain. The four steps are described below, based on "Mastering Bitcoin" (Antonopoulos, 2014) and "Bitcoin whitepaper" (Nakamoto, 2008).

Generating keys and creating a transaction

Each agent that wants to interact in the blockchain has a private key which only him (or the application managing the access e.g. a digital wallet in the example of the bitcoin blockchain) can see and generates the corresponding public key. This public key is the address of this agent (after a one-way function cf. figure 16) and is visible by all peers in the network, allowing other agents to interact with it. In the case of a payment transaction, this is the address to send the funds to. The public key (PuK) is generated from the private key (PrK) but the private key cannot be found by knowing the public key. It is a one-way relationship that is impossible to reverse.

When Alice wants to make a transaction with Bob, she signs the transaction with her private key PrK. This corresponds to encrypting the transaction (T) in a message (M). Anybody in the network can easily verify that the message M is in fact the encryption of transaction T by

the private key PrK corresponding to the agent's known public key PuK (address), without revealing the private key. Verification and generation use elliptic curve cryptography.

$$PuK = generate(PrK)$$
$$M = encrypt(T, PrK)$$
$$Verify(PuK, T, M) = True\ or\ False$$

Example of bitcoin address = Hash(PuK) = 1Cdid9KFAaatwczBwBttQcwXYCpvK8h7FK

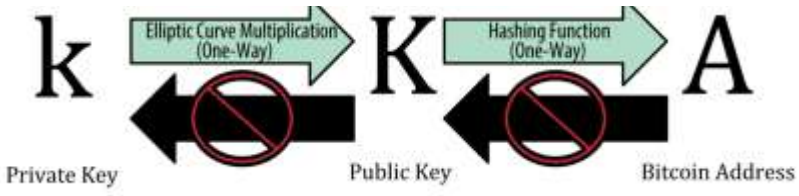

Figure 15: One-way transformations, from private key to address (Antonopoulos, 2014)

Broadcasting the transaction to the network

The combination M, T, PuK is then broadcasted by the blockchain *client* (i.e. wallet in the case of a cryptocurrency payment) to the network of nodes. It is important to notice that it is a peer-to-peer distributed network, meaning that each node will connect with some other nodes and the message will thus be distributed quickly to several nodes which all check the validity of the signature via the verify function. The process of broadcasting unseen transactions is called "flooding." The *full nodes* are the agents that have a copy of the blockchain, ensure the routing to share information with other nodes while lightweight nodes only download the headers of block (see later), only to validate the authenticity of transactions. The transaction being propagated to the network does not mean yet that it has been added to the blockchain.

Adding transaction to the block & Linking the new block to the blockchain

The encrypted transactions are first added to a pool of transactions and selected by miners, full nodes with specialized hardware, who will add these transactions into blocks in a process called mining. The transactions are picked by miners, generally (but not always) based on the transaction fees given by the users i.e. the higher the fees, the more chances to be selected for a transaction. Miners verify transactions and hash them. Hashing corresponds to the unilateral action of taking any sequence of information and generating a standardized output which does not allow to reconstruct the input only based on this output but is highly improbable to reproduce with any other sets of inputs. A typical hash function is SHA-256 which can transform any input in a 256-bit long output.

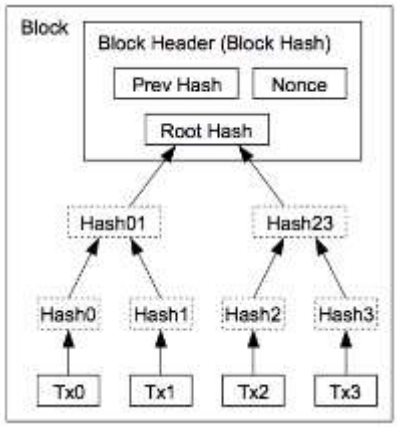

Figure 16: Block structure & Merkle tree. (Nakamoto, 2008)

Hashes of transactions are summed two-by-two in successive steps (see schema below), a structure called *Merkle tree* which ultimately gives one output, the *Merkle root*, which is a summary of all hashes of all transactions, meaning that if any transaction is modified, this Merkle root changes completely.

Each block is represented by a unique hash containing all its information, including the hash of the previous block (parent block) in the chain. By going backwards in the chain following this chain of hashes, we finally arrive at the first block produced (in the timeline) which is called the *genesis block*. Because each block has the hash of its parent block, if any blocks are subject to any modification (even minor), all children hashes are changed, and the blocks need to be recalculated.

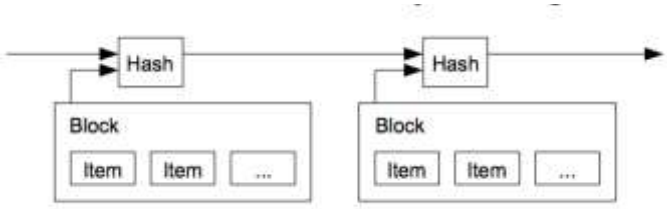

Figure 17: A chain of hashes connecting blocks creates immutability (Nakamoto, 2008)

Without going too much into details, calculating blocks consumes a significant amount of energy, which costs are assumed to be higher (taking into account the risks related to an operation which may fail) than the returns that a malicious actor can make by changing transactions, ensuring the security of the network.

In practice, calculating blocks means finding a random number (called "nonce") which when hashed with the block hash, results in a hash starting with a given number of zeros. This number of zeros is called the difficulty (after some mathematical operations) and is adjusted dynamically to make sure that blocks are created on regular time intervals of 10 minutes

independently of the hashing power invested. In exchange for this proof-of-work, miners are rewarded with newly minted cryptocurrencies (bitcoins in the case of the bitcoin protocol).

Each block contains a header containing metadata (data about the data within the corresponding block) and a body with all transactions. In the block header are contained the hash from the previous block, the information related to the mining part, and the Merkle root of this block as explained previously.

Example of a block hash (the number of 0 in bold corresponds to the difficulty to mine coins)
000000000019d6689c085ae165831e934ff763ae46a2a6c172b3f1b60a8ce26f

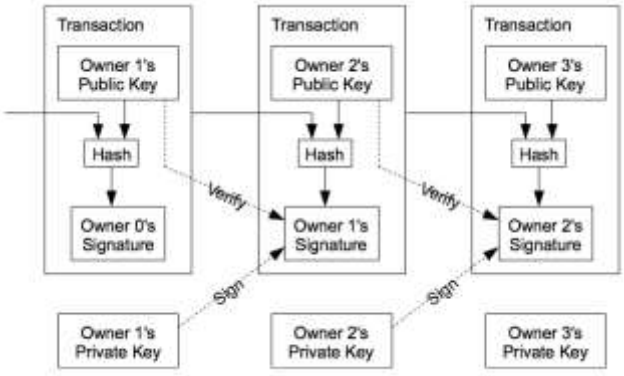

Figure 18: Blockchain architecture (Nakamoto, 2008)

As a conclusion, we see that because each block is connected to the previous one, no modification is possible without the network noticing it (and rejecting if necessary). The state of the network is quickly spread to all participants so that all stakeholders have a common view of the ledger. We can now explore the values previously mentioned and justify them with high level explanations.

The blockchain architecture as explained above may vary considerably depending on the blockchain settings. The Bitcoin blockchain is a *permissionless* and *public* blockchain, meaning that everybody can read and write in the blockchain, but other variants exist.

This data-structure potentially adds the following values to current systems:

- *Decentralization*: transactions are processed, validated, and stored by a large number of nodes which are coordinated by a consensus protocol instead of by one or several actors. There is no single actor take can control all decisions. Governance is ensured by consensus.

- *Censorship-resistance*: current blockchain such as the Bitcoin one verifies the validity of the transaction, but not the content in itself. As long as the transaction matches with the

protocol rules (which are open-source) the transaction will be added to the blockchain and cannot be refused by a central-authority.

- *Borderless*: everybody is in theory able to run a full node (in practice some limitations exist such as available storage or the need for understanding the Bitcoin protocol to some extent). It should therefore not depend on geographic location.

- *Immutability/Security:* thanks to the chain of hashes, it is impossible to change a transaction without obtaining consensus about it. In order to force a consensus about false transactions, miners would need to have more than 50% of the hash rate, which means they need to provide more than half of the computation and thus significant resources, making it in principle too costly and risky to attack the network if there is enough hash power deployed by other stakeholders.

- *Transparency:* on public (more on the distinction private vs public in section XX) blockchains, anybody can access the transactions and thus see which addresses are making the transactions. It is thus possible to follow the transfers of assets between public addresses. Public blockchains also generally have the code in a repository online (typically on GitHub) which everybody can consult and contribute to.

- *Efficiency/auditability:* as all transactions are stored in a common shared distributed database, it is easy for all actors to check transactions and in particular to use these transactions as proofs. In fact, as they have been digitally signed by a unique private key, assuming that it has not been stolen, one can easily verify agreements with other parties. In the context of data sharing, it will be crucial to record that data providers give their consent for the use of their data, and in the other direction data supplier must write in this database the terms and conditions, for example the purpose of the data usage. It is then much easier to take legal actions for both parties, if other parties appear to show dishonest behaviors.

5.3 Blockchains in data marketplaces

According to Karafiloski et al. (2017), blockchain holds many promises for data by giving more control to the data providers, increasing transparency, for storing data in a distributed manner, for user authentication. Ramsunder et al. (2018) have formalized token-based data markets on a mathematical level, including the operations that are made possible on data structure and their associated parameters, as well as potential attack vectors. Roman & Stefano (2016) have proposed a reference blockchain-based architecture for exchanging data used for credit scoring. The emergence of these blockchain-based data application increases the relevance of blockchain for building our sensor data marketplace. However, many blockchain applications and properties are not relevant for our design. In the following section, we list and describe the important ones.

5.4 Relevant concepts for the design phase

5.4.1 Decentralized permission system with off-chain storage

Zyskind et al. (2015) have proposed an access-control management system for data providers to easily give or revoke access to the data they provide. The access policies are stored on the blockchain, while the blockchain connects to the data via a distributed hash table (DHT) (Maymounkov & Mazieres, 2002). The blockchains therefore acts as a mediator between the data provider, the data user and the data. To interact with the blockchains, data providers can submit transactions that can be of two different kinds: *Access transactions* (only by the data provider) for access control management and *data transaction* for data storage and retrieval. Data users can request data by submitting a data transaction to the blockchain. *Nodes* maintain the network, in return for incentives. By maintaining the network, we mean that they verify transactions and make the proof-of-work to secure the network and adding new blocks to the chain, with some newly minted tokens.

When there is a request for accessing data from the service and that the data provider is willing to share the data, he grants access via an access transaction. He also sends the data that become encrypted by a shared key i.e. a combination of his public key and the service's public key. The data is directed by the blockchain to an off-chain database and hashed. The hash is recorded on the blockchain and acts as a pointer to the data via a distributed hash table. This table is a key-value table with the keys being the hashes and the value being the corresponding encrypted data in the off-chain storage system. Therefore, the data user can query the data by using the pointer and proving that he has access to the data as the blockchain automatically matches the service set of keys with the latest access transaction from the user. To grant new access or stop current accesses, the user can create a new access transaction. It enables the system to reach a very granular level, meaning that the user is able to decide which specific business can access the data.

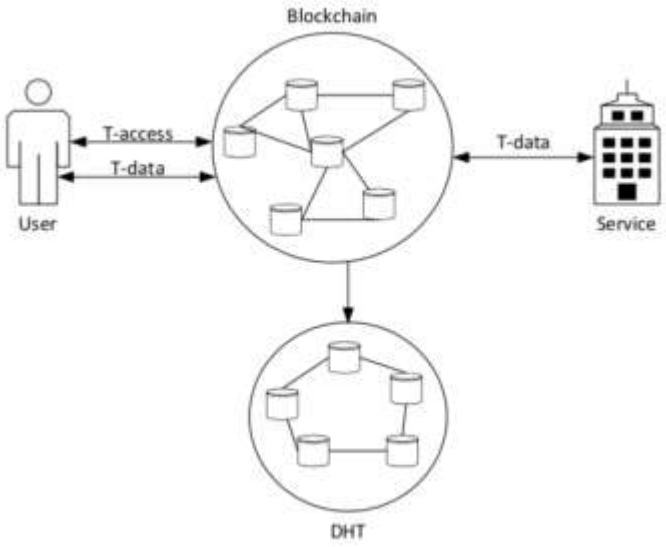

Figure 19: Decentralized permission system (Karafiloski et al., 2017)

Note: the "User" in the figure corresponds to what we have called data provider while "Service" refers to our data user. These seem counterintuitive and result from different notations between our work and the paper from Karafiloski et al.

Based on the analysis of the paper from Karafiloski et al. (2017), there is one clarification question remaining: what happens if there is more than one user that wants to access the data? The paper suggests that data are encrypted via a shared key resulting from the user and the provider. We are not sure if it implies that for every new user, it would mean duplicating the data and encrypting it again with a new shared key (for the provider and the new user). This would be highly consuming in terms of storage. We rather believe what the authors are trying to achieve is a system where every time there is a new user, a new shared key is created, and the data are re-encrypted, such that keys of all users and of the provider can decrypt the same
encrypted dataset.

5.4.2 Smart contracts

Smart contracts are self-enforceable programmable contracts (potentially stored on the blockchain) that respect and operate according to a predefined logic and set of business rules (Buterin, 2014; Szabo, 1997). These smart contracts are the core of decentralized applications and allow the inclusion of conditions when exchanging digital assets. For instance, smart contracts can be used to keep funds in escrow until certain time delays are met.

5.4.3 Token curated data

In their whitepaper and interview, Van der Veer et al. (2017) describe token-curated data: a way to increase trust in the quality of a dataset based on the "put your money where your mouth is" principle. It is a way to circumvent the "cold start problem" typical of e-commerce. E-commerce platforms usually work with ratings and therefore new comers face difficulties in proving the quality of the product/services they are offerings since they do not have any ratings yet.

In a token curated market (Ramsundar et al., 2018; Ocean Protocol Foundation, 2018), providers bet on their service/product by putting tokens at stake in escrow using a smart contract. Other stakeholders in the system see the amount of data that is put at stake and perceive it as an indicator of the quality since if the terms of the contract are not met, the data providers will lose its tokens. For instance, in our data marketplace, if the dataset does not include what it was supposed to (e.g. according to a description), the tokens are lost. The tokens are not only lost but given to the stakeholders who bet *against* the provider, by themselves putting tokens at stake in the escrow. It is necessary that they also put tokens at stake in order to avoid sybil attacks (Douceur, 2002) i.e. many agents betting randomly against many providers in order to maximize their chance of winning some tokens. In addition, other stakeholders can also bet by putting their tokens at stake. For instance, if they have bought a good or service and have judged the quality as matching the description, they can confirm. Following the same logic, they can also bet against. At some point when too many opinions are conflicting, a stakeholder can ask for a *conflict resolution* meaning that some mechanisms in the system will verify who is right and give all the tokens in escrow to the side that is right. This conflict resolution part is the most difficult part of the process since it is not always possible to resolve it via code or via the stakeholders directly involved. It is sometimes necessary to include a third-party or "*oracle* "which acts as a judge. In this case, we notice that the system may lose in decentralization. The decision could also be crowdsourced, but this may result in inefficiencies. It all depends on the context e.g. the participants in the network. In a permissionless blockchain environment there are too many participants and they cannot be trusted. An oracle is needed. In private settings, generally participants were accepted in the network because they are trusted, and a vote makes sense. A vote does not imply that all participants need to be online at the vote moment. They could also delegate their vote to some representatives who would take the decisions.

5.4.4 Other enabling technologies

Homomorphic encryption

Homomorphic encryption (Gentry & Bonneh, 2009) is an encryption mechanism that preserves the same form between the message and the cipher-text (the encrypted message). This allows the data users to perform computations such as training their predictive models, without decrypting the data. This differs from traditional practices that require users to first decrypt the data before being able to perform any meaningful operation. Although this technique has already been introduced several years ago, its practical use is still questioned by some cryptographers (Interview 5) since the mathematical operations that can be applied

on the cipher text are limited. As a conclusion, it is possible, in limited applications only currently, for data providers to monetize their data to benefit data users without even having to decrypt it, and therefore by keeping a high level of data protection.

Chapter 6: Design of a decentralized sensor data marketplace

6.1 Introduction

In this chapter, the context-aware system design previously introduced is applied to build a decentralized data marketplace system that can support sensor data exchanges with businesses.

In Section 6.2, we elaborate on the requirements that the data marketplace must respect. These are basic features that are expected, independently of the context. In Section 6.3, we start with giving a summary of the design method. Section 6.4 is about getting insights into the context by defining the *foci* and situations that restrict the foci, as well as the context elements. In section 6.5, sensors and adaptors are defined. In section 6.6, we establish the rules for reasoning from the information received by sensors to the update of adaptors. All information is then summarized in section 6.7 before discussing and assessing the design in the following chapter.

6.2 Data marketplace requirements

This section aims at providing the base for the data marketplace design. The elements resulting from the requirements are not context-aware and are therefore provided before applying the design method.

The main function of a marketplace is to bring together demand and supply for some goods. Therefore, there needs to be a physical or online place where both sides can gather. As the commodities traded here are data and not physical goods, it is obvious that the marketplace must be online. A domain and webpage are therefore required to allow providers and users to navigate, upload their data, browse the data catalogue, and other actions that they need to take depending on their need.

The user-experience is crucial if we want to scale sufficiently the marketplace. The front-end needs to be very user-friendly, and the back-end must operate fast enough to not create latencies on the user side. For the front-end (what is presented to users), a basic requirement is to have a dashboard on this webpage, where providers and users can log in, see the data they bought, and look for others. When identity matters, and it often does as stakeholders want transparency (interview 3), ID authentication systems should be used, such as digiD or uPort. Concerning the back-end, speed and security are basic requirements. However, we cannot detail more on this since it depends on the technical implementation of the marketplace. For instance, the database choice will impact differently data retrieval speed and security.

As data are bought and sold, the marketplace must also connect to a payment gateway. Payment solutions could include credit cards and cryptocurrencies. It is important to remind the reader that at no point in time has the platform developer access or control over these payments. These are done externally, via the payment provider, who notifies the system that the payment has been processed. This means that in addition to having no control over the data, the platform owner also has no control over money.

Finally, the website should be combined with a mobile application for data providers. This would allow them to easily give access to data when they get notified. We believe that it is less relevant for the data user, since manipulating large datasets is something that is not done on a smartphone anyway and therefore the user needs to be on a computer/laptop.

6.3 Context-aware method summary

6.3.1 Method overview

The context-method presented in van Engelenburg et al. (2018) includes three main steps: (1) Understanding the context, (2) determining components that can sense the context and adapt, and (3) determining the reasoning rules that govern the adaptors, based on the context information sensed. Step 1 & 2 are decomposed in several sub-steps, as illustrated on Figure 20.

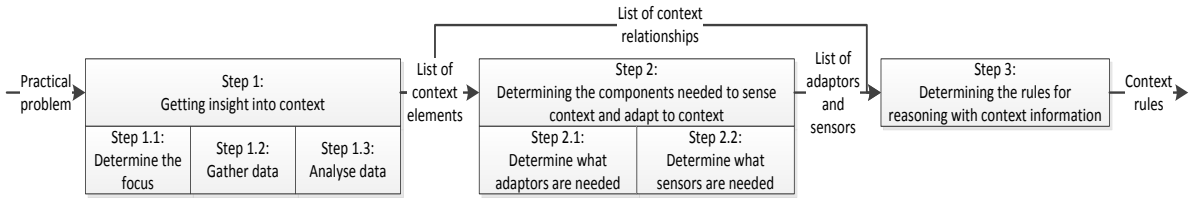

Figure 20: Overview of the method (van Engelenburg et al., 2018)

We list below the definitions of the technical terms present on the diagram. These terms will be defined again in the next section and explained along with the process. We provide only the semantics definition from the paper. However, the paper goes deeper in the logic and provides also syntax definitions for these terms. Without the context of the paper, it may be difficult to understand these definitions as they are quite abstract. Therefore, after providing each definition in Section 6.3.3, we provide an explanation with our own words, and illustrate with a running example of our decentralized data marketplace.

The *focus of a designer is the relationship between entities that the designer needs to have a certain value to reach their system* goal (p.10)

A context relationship is a relationship between a focus and a minimal set of attribute relationships, where in each situation where these attribute relationships have the same truth value, the focus has the same truth value. In addition, it should be possible that there exist situations in which these attribute relationships are not true. We say that these situations restrict the focus. (p. 12)

A *context element is an object relationship which is part of a set of object relationships that have a context relationship with the focus* (p. 14)

The context of a focus is the set of all its context elements. (p. 14)

6.3.2 Schematic literals and predicates

The context rules introduced on Figure 20 and developed during the design phase are written using *schematic literals*. A literal includes variables in its terms, such as $isProvidedWith(User, Information)$. It allows the designer to translate the object relationships into logic that can be understood and executed by a machine. We also introduce the notion of *predicate* as it will be important for the understanding of the reasoning rules. A predicate is a function f: A $\rightarrow$ {0,1}, that is, for all x $\in$ A, f(x) $\in$ {0,1}. This function takes any input from the domain given A, and as an output gives a Boolean value (mathematically formulated as 0 or 1 in the definition). An example of predicate is $isProvidedWith(User, Information)$. It takes as input any object from the set of users and information belonging to the domain, and as an output gives True/False i.e. it tells us whether a specific piece of information is shared with a specific user or not.

6.3.3 Method summary

In order to interact with the environment, it is first necessary to get insights into the context. This process is divided in three parts. The first step is to define the focus from the problem as initially introduced in the problem statement. The *focus of a designer is the relationship between entities that the designer needs to have a certain value to reach their system* goal (van Engelenburg et al., 2018, p.10). In other words, the designer has a specific objective that he is trying to achieve with the design, called the *goal*. To reach this goal, the designer delivers attention to the elements that he knows (generally based on interviews with stakeholders) will contribute to it. These elements can have different states or values, and the designer wants these to take the specific combination of values that the designer concluded will enable the goal. The *focus* is a more formal description of these elements, which are relationships between entities.

For instance, if a designer wants to build a marketplace (goal), he needs to have both buyers and sellers participating (articulation of the definition of a marketplace). The focus is

therefore this willingness to participate, which can more accurately be described as a relationship between entities (e.g. buyer, seller, level of willingness). The designer wants this relationship to have a certain value: willing to participate (versus unwilling to participate).

Once the foci have been defined, the designer collects information related to these foci, or more accurately about *situations* that impact the foci. *A situation is a state of the world, determined by a possible combination of attribute relationships* (van Engelenburg et al., 2018, p.10). A situation is therefore a static representation of the stakeholders and other objects, and how to they interact based on which attributes do they have. For instance, a situation could be the presence of a data user, a data provider and a dataset uploaded on the marketplace by the provider and that is interesting for the data user.

The data gathering step will be done using a literature review and stakeholders interviews. Once enough situations have been defined, the main object relationships in the situation must be highlighted.

At this point, it is necessary to reduce the scope to keep the research feasible by prioritizing the situations and selecting only the most appropriate ones. The priority can be defined according to various criteria such as the number of occurrences where a situation is mentioned by an interviewee, or by asking them to rank the situations. It could also be up to the designers to decide which situations should be selected for further steps, as the design has a global view across all interviews, complemented by the literature review. Finally, the context elements of each focus can be listed. Explaining context elements with our words: there are objects that have context relationships with the focus, meaning that they influence the values that the focus can take. These objects have relationships and form a set of object relationships. Each of these is a context element. Readers that intuitively thought that an element was an object must understand that it is actually a relationship between objects.

For instance, assuming the focus is still the willingness to participate. Then privacy is a context element, since it is a relationship between objects (e.g. buyer, seller, other stakeholders, transaction, level of privacy) and as it has a clear impact on the focus (it is easy to imagine situations where an individual would not want his peers to know that he is buying some goods).

Once the context is defined, the second part consists of finding the sensors that will detect changes in the environment. With the right sensors, the system is receptive to stimuli if they are part of the context. After an event has been reported by the sensors, the system must react according to predefined rules. Two additional types of components are needed. The first category includes adaptors which will enable a reaction of the system to the change in the context by updating some parts of the system. The final component that is required is the brain of the system, the bridge between sensors and adaptors: reasoning components. Based on the inputs received by the sensors, reasoning elements give different outputs to adaptors to match with the context according to the designer's definition. Figure 10 summarizes the method using a diagram. The context-aware system senses changes in the part of the

environment that belongs to context and adapt to these changes (according to defined reasoning rules).

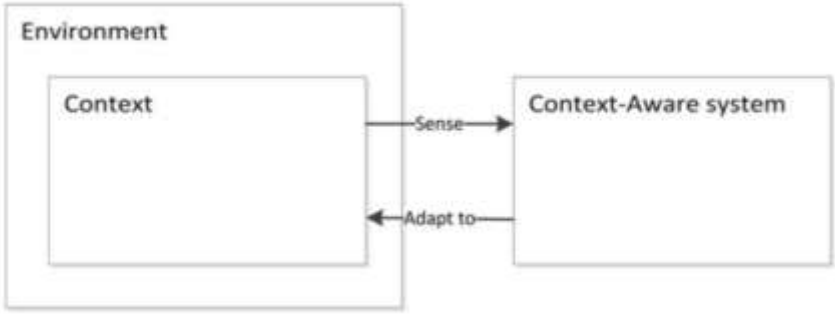

Figure 21: Context-aware system (van Engelenburg et al., 2018)

6.4 Getting insights into the context

6.4.1 Defining the foci

In the first step, the objective is to learn about the context that is relevant for the data marketplace. This can only be done after defining successively the problem, the goal of the designer, and the state of the world when the goal is reached. The problem has already been discussed extensively and has been described as the unwillingness of stakeholders to participate in the data marketplaces, mainly because of the presence of intermediaries in infrastructure for exchanging data between sensors owners and data users.

Goal

The goal of the designer is therefore to enable data flows between sensor owners and data users in an efficient and secure way, as much peer-to-peer as possible, with sufficient incentives for data providers to participate in the data marketplace.

State of the world

When the goal is reached, the state of the world is that data marketplaces can be trusted by the users since they trust the technology, or more accurately the trust is created by the technology forcing stakeholders to behave correctly and by increasing transparency. As a consequence, a large amount of data is produced and exchanged to create more value per data as different stakeholders are using these. Many more data-driven decisions and research can

be undertaken as the infrastructure is providing a data market on a large scale. Sensor owners have control over their data and are willing to share, and data users are also participating by acquiring relevant data.

As a second step in getting insight into the context, based on the problem definition, goal, state of the world, general knowledge about marketplaces, desk research, and interviews with stakeholders, the focus points can be defined.

Foci

The two first foci are necessary conditions for having marketplaces which in essence need both a supply and demand side to exist. Therefore, both data buyers and sellers need to be willing to participate in the marketplace. They may be willing to participate with different intensity (low, middle, high). The two foci can be formalized and written as predicates (defined in 6.3.2). As a reminder, a predicate is a function f: $A \rightarrow \{0,1\} \mid \forall x \in A, f(x) \in \{0,1\}$.

Focus 1: Sensors owners (entity) are willing to participate (relationship) in the marketplace (entity). This is a relationship between these sensor owners, the data buyers (entity), a flow of data (entity) and a level of willingness (entity).

$$WillingnessToParticipate(SensorOwner, DataBuyer, DataFlow, levelOfWillingness)$$

Focus 2: Data buyers (entity) are willingness to participate (relationships) in the marketplace (entity). Relationship between these sensor owners (entity), the data buyers (entity), a flow of data (entity), and a level of willingness (entity).

$$WillingnessToParticipate(DataBuyer, SensorOwner, DataFlow, levelOfWillingness)$$

The predicate above expresses the willingness to participate of the first entity within the brackets (respectively the sensor owners and the data buyers). The second argument can be understood as the stakeholders that the subject will interact with.

In order to indicate that a proposition does not hold, we use the negation of the predicate. In this case, if the data buyer is not willing to participate, we write:

$$\neg WillingnessToParticipate(DataBuyer, SensorOwner, DataFlow, levelOfWillingness)$$

As we already know, the main points impacting these foci are the centralization of data exchange-related processes that we concluded in section 3.4 by saying that the lack of trust is a main reason for not participating in a data marketplace. In addition, the notion of a large-scale system appears in the state of the world definition: “the system needs to be able to sustain large scale data exchanges”. Scalability is one of the main issues in blockchain-based

architectures (Xu et al., 2017). However, this is not something that changes with the context, it is always true. In other words, the scalability issues do not require a different solution in different situations. It is therefore not a third focus but part of the basic requirements of the design. However, as public blockchain scalability is still an unresolved and very complex problem, we will not develop this further. It is left for future research when more scalability knowledge will be available.

6.4.2 Collecting data

The scope has been reduced to two main foci. It is therefore known which direction the data collection should take in order to discover more about situations that can influence the foci. Data on the situations restricting the foci are collected using literature review and interviews. The arguments have already been presented in Chapter 4 in the knowledge base. However, we will formalize them in relation to the context by articulating how do they influence the foci. More specifically, to be able to formalize the context relationships, we use the criterion suggested in the method:

"Criterion: An object relationship has a context relationship with a focus, if and only if, it is part of a minimal set of object relationships, such that there are values for each of the object relationships in the set, such that in each situation where they have these values, the value of the focus is the same." (van Engelenburg et al., 2018; p. 15)

Focus 1: Sensors owners' willingness to participate in the decentralized data marketplace".

$$WillingnessToParticipate(SensorOwner, DataBuyer, DataFlow, levelOfWillingness)$$

Based on the data collected and the established criterion, we have defined six high-level descriptions of situations restricting the first focus. However, we do not detail all descriptions since we will apply the complete design procedure to two only. Only these two will be discussed exhaustively, and only these two will be presented with the right syntax of a situation as defined in 6.3.3. Therefore, we nuance the vocabulary: first we discuss "high-level descriptions," and then when going on a lower level, we will use "situations". We may also omit the "high-level" term for readiness purpose.

The first high-level description was explained in the article of Roman & Stefano (2016), which suggested that "privacy, security, and control of data increase the willingness [of data providers] to share their data". Sensor owners produce data that have an interest for different stakeholders. However, they may not be willing to share the data with all of them. Specifically, there is a high chance that the data they produce also interest the competition, since the data may be directly related to the product or service that the business is offering. They may therefore prefer to restrict the access from competitors. Another possibility is that the sensor data may reveal some insights that should be kept hidden from regulatory bodies. For instance, the autonomous vehicles speed data may reveal excesses in speed. These data

are still useful for training algorithms, but the sensor owner does not want to be fined for breaking the law. As a conclusion, if one of these stakeholders is able to acquire the dataset by buying it on the data marketplace, the sensor owner will not want to participate. We defined the first description restricting the focus to *sensors owners do not want to participate:*

High-level description 1.1: The data is sensitive for the data owner to some businesses. The data is shared with the business the data is sensitive to.

The second description was suggested in the third-interview, by a consultant in a research-oriented technology firm. The interviewee emphasized the "mess" that firms are currently facing with the General Data Protection Regulation. He related that "firms are reluctant to share data because of fear". In our case, the data produced are not personal data and are therefore not subject to the GDPR, despite that boundary not being always clear and therefore a sensor owner may not be sure about on which sides does it stand. However, it illustrates the broader legal risk implications that businesses may fear, which could also cover copyrights, doubts about the consequences of sharing inaccurate data. For instance, what happens we discover that sensor owner A has provided self-driving car data to company B and this appears to incur a loss for company B? Where does sensor owner A stands in terms of legal responsibility? Shall A offer a compensation to B or is A's responsibility limited to a moral or social one? These questions are examples that could come through the sensor owner's mind and restrict him to not participate in the data marketplace. We therefore formulate the second description:

High-level description 1.2: The data are shared. Sharing the data has legal consequences that the sensor owner may not know about.

The third description was extracted from the interview with a data provider, sharing energy sensor data. They mentioned that one of the main points that would discourage them from using a data marketplace is providing their data to a third party that they cannot trust, assuming the data are proprietary and have some value that they want to benefit from. In other words, assuming an open-data context, they would be willing to give access to the platform provider since the data are already available. However, if they expect the data to be monetized on the platform, then if they do not trust the market place provider, knowing that the marketplace provider can access their data could restrict their willingness to participate. When asked about more details about which actions from the marketplace provider would discourage them, they mentioned "sharing with other third-party that we do not want to share with" (which relates and thus confirms description 1.1). However, they also state that the marketplace provider could share the data with businesses that they *are* willing to share with, via private secret contracts. This would take away the benefits of the sensor owner. The marketplace provider could also change the datasets and for instance reduce its quality, if there is some sort of conflict of interest and that the platform is not neutral. This is particularly relevant for the self-driving cars development since several companies behind the autonomous vehicles software developments also host cloud services (e.g. Google,

Microsoft) which can also host data marketplaces (e.g. Google Cloud platform, Microsoft Azure Data marketplace). The third description is proposed, leading to sensor owners being willing to participate:

High-level description 1.3: The data is proprietary and has value. The sensor owner does not trust the platform provider. The marketplace provider cannot access the data.

The fourth point also discusses transparency, but more downstream in the data flow: when the data have been acquired by users. Karafiloski & Mishev (2017) have criticized the lack of transparency about how sensor data are used. In their words: *"[…] users still don't have a clear preview on which precise data is collected and for what purpose. Users lose total control of what happens with the data."*. We have mentioned previously that data are important and can create significant impact. We have mainly discussed the positive impact, but there are cases where some users may use data unethically (according to the data owner's ethics). The lack of transparency restricts the sensor owner to *not willing to participate*.

High-level description 1.4: Sensors owners do not know who is using their data and for what purpose. Sensor owners would like to have more transparency about who has their data and how are they being used.

When proceeding to the interviews and asking about the marketplace provider, one point appeared to be generalizable to more than the marketplace provider: sensors owners do not want to lose ownership of their data. The interviewee used to be a centralized data exchange, before moving to a decentralized version. One of the main feedback they receive from customers is the sentiment that once uploaded, the data are not really the property of the data sensor anymore. As the dataset is duplicated they are afraid that other parties claim that they have produced it. Using their words: *"Transparency and ownership of data are important factors that complement the need for privacy and security of the data. Not having these characteristics fully operational was one of the biggest barriers before for our centralized exchange. For every single dataset, we should always know who the author is, almost like with citation in the academic field. Some are not satisfied with only a financial reward, which they may also lose if they are not officially owners anymore."* We therefore deduce that there is a need for traceability of a dataset to its owner. As we will see when analyzing the second focus, the upstream transparency is also expected from the data user side.

Considering the context-awareness relevance, it can be argued that sensor owners always want more transparency, and therefore that it does not depend on the context. It is true that most of the time sensor owners will want more transparency. However, it is not always true e.g. if a sensor owner has the certainty that a dataset cannot lead to any unethical consequence, or simply that the sensor owner is not concerned about whom the data are shared with, then he does not require more transparency. This is especially true in a sensor data market, where the sensor owners are measuring (physical/technical) phenomena and are trying to monetize these measurement, instead of selling personal data. Nevertheless, this

does not imply that the designer of this system believes that it cannot lead to unethical consequences. The wrong sensor data in the wrong hand could definitely have a negative impact independently of whether the data provider wants more transparency or not. The attribution of responsibilities, not only legal but also moral ones, is important to consider and will be introduced in the discussion section (Chapter 8).

High-level description 1.5: *Sensors owners lose ownership attribution for their data.*

The sixth description is a more practical one, as Chen & Xue (2017, p. 5) have mentioned that "*organizations have rich data resources and also would like to share or sell their data. However, they usually do not have enough IT engineers and it is very hard for them to provide these services."* When discussing this with interviewer 2 and 3, and when analyzing IT resources from a corporate perspective, we observed that this lack of IT engineers focusing on the data sharing shows that the incentives to share data may not be large enough. This can imply that the financial benefits or brand recognition are not worth giving up IT resources. Depending on the incentives, on the firm's resources, and on the user-experience related to uploading datasets or connecting data streams to the data marketplace, data owners will or will not participate in the data marketplace.

High-level description 1.6: Sensor owners have limited IT resources for uploading data to marketplace. There is not enough incentive for taking these initiatives.

This description is also only considered on a high level as it is not selected for the next step. Incentives could be of different kinds, we did not specify which ones, but potential ones include financial or merit (e.g. visibility).

The table below summarizes the six descriptions that restrict the first focus, as well as the sources supporting the claims. By restricting the focus, we mean that based upon the values of the objects in the descriptions, the focus takes one or another value.

Table 2: Situations restricting the focus 1 (sensor owners)

Restriction to focus	Description	Support
Description of how the focus is restricted	Description of the high-level description in which the focus is restricted	Reference to a data source
Sensor owners do not want to participate	The data is sensitive for the data owner to some businesses. The data is shared with the business the data is sensitive to.	"*Privacy, security and control of data increase the willingness [of data providers] to share their data."* - Roman & Stefano (2016, p. 6)
Sensor owners do not	The data are shared. Sharing the	"*There is currently a huge*

want to participate	data may have legal consequences that the sensor owner does not know about.	*mess around the GDPR within firms. They are reluctant to share […] because of fear."* – Interview 3
Sensor owners want to participate	The data is proprietary and has value. The sensor owner does not trust the platform provider. The marketplace provider cannot access the data.	*"As we do not trust the cloud provider for accessing our data, we would prefer a platform with a proof that the owner is not accessing and using our work for other purposes."* – Interview 4
Sensor owners do not want to participate	Sensors owners do not know who is using their data and for what purpose. Sensor owners would like to have more transparency about who has their data and how are they being used.	*In critics of current data marketplaces section: "[…] users still don't have a clear preview on which precise data is collected and for what purpose. Users lose total control of what happens with the data."* - Karafiloski & Mishev (2017, p. 7)
Sensor owners do not want to participate	Sensors owners lose ownership attribution for their data.	*"Transparency and ownership of data are important factors that complement the need for privacy and security of the data. Not having these characteristics fully operational was one of the biggest barriers before for our centralized exchange. […] "* – Interview 1
Sensor owners do not want to participate	Sensor owners have limited IT resources for uploading data to marketplace. There is not enough incentive for taking these initiatives.	*"These organizations have rich data resources and also would like to share or sell their data. However, they usually do not have enough IT engineers and it is very hard for them to provide these services."* – Chen & Xue

		(2017, p. 5). Also discussed with interviewees 2 and 3.

After having collected and analyzed data about the first focus, we now study the second one, on the data demand side of the marketplace. As a reminder, the second focus is:

Focus 2: Data users' willingness to participate in the decentralized data marketplace.

Following the interviews and literature review, there are four main situations that can be listed. All these restrict the data users' willingness to participate in the decentralized data marketplace to the value *not willing to participate*.

One of the main reasons as underlined by the various interviewers for the demand side to not participate in the marketplace is the risk of not having quality data. Quality data is quite a vague term and need to have further specifications. The data acquired by the data users are considered as being of a higher quality if the number of data points and the data format (e.g. JSON, xml) are the same as stipulated in the data description: if the methods used for collecting the data are scientifically correct, if the sensors match specifications, and if they give an accurate measurement of the phenomenon they are trying to measure. As an example, assuming that the data buyer is looking for an accurate measurement of the air quality in Belgium, he needs a sufficient amount of data points coming from sensors spread around the country to be able to derive a conclusion at a national scale. If a dataset description claims to include such measurements but which in reality only has a few data points (or many data points coming from only a few sensors), the data will be inaccurate. After a bad experience, data users may not be willing to participate in the marketplace anymore.

Interviewees were concerned about the data quality perception by the user. Since users cannot rely on an intermediary to curate or at least check the data, they may not trust the data supply side. Without a proper way to convey confidence of a data quality, it is unlikely that they will participate in the marketplace. However, the expectations of the users in terms of quality can vary and can therefore change their threshold such that they accept a lower quality. As an example, if they have a data scientist available to clean the data, they may be willing to lower the threshold. Also, there are indicators in the context that will influence their perception of the data set quality (e.g. when it was posted, by whom, what does the description of the dataset says…). All these elements can influence the participation of the demand side, and therefore put the whole marketplace at risk, since one side cannot exist without the other.

Situation 2.1: Upon query, data are proposed to the user. Data quality does not match with the requirements of the user.

The second proposition is straightforward: the data must be useful. If data users have specific needs but cannot find relevant information in the marketplace, they will not participate. The third interviewee has formulated this requirement by saying that "*As a data user, I need to be*

able to find sensor data that actually benefit my business by for example improving my production predictions". Data needs may be very specific. For instance, households energy consumption shows a high seasonality, as the consumption is much higher during winter than summer. If a business is trying to optimize its grid production in the winter but has only summer-related data points available on the data marketplace, he will not participate. However, if the information is relevant, he will be willing to participate. Therefore, we formulate a situation restricting the focus to willing to participate.

Situation 2.2: Data is useful i.e. there is perceived added value for a stakeholder resulting in acquiring it.

The third situation is referred by interviewee 4 as "data scarcity". With this term, he suggested that the marketplace would only effectively work if the data are not available elsewhere at a lower cost. However, after analyzing this statement, we believe that it is more accurate to specify that the lower cost should not only be the monetary value of the dataset, but also the transaction costs associated to searching the data in other sources and proceeding to the transaction (with a potential risk if the other source is not secure). As observed from the literature review about marketplace, the main function is to bring stakeholders together to remove inefficiencies related to one-to-one exchanges. Now, there may be another data marketplace offering better price and a better service which would then attract the data user. It is therefore important to sense the context to see what the other available opportunities are, and potentially price dynamically. However, we mentioned earlier that discussing economics and pricing mechanisms is not part of this research. Therefore, we just mention the economical dimension of this situation but will not investigate it further.

Situation 2.3: Data buyers have to pay for getting the data on the marketplace. Data is available elsewhere at a lower cost.

Finally, the last proposition is about the transactional privacy. Interviewee 3 said that "*Privacy is also important on the data user side, since it may give insights into their strategic plans"*. By "may give insights into their strategic plan", it is meant that the data acquired by a business on the marketplace may reflect some of their future actions. For instance, if business A acquires self-driving car data from several actors on the marketplace, it is likely that they are considering entering the autonomous vehicles industry. It is not sure, as they could be also using data for other purposes (e.g. complementary product or service), but it represents a risk that businesses may not be willing to take. Therefore, it will depend on to what extent do they want to take the risk, which depends on the sensitivity level of the data acquisition. Following from this, we can formulate the last situation restriction the second focus to *data buyers are not willing to participate to the data marketplace*.

Situation 2.4: The transaction is sensitive for data users to the competition. The competition can see the transaction.

Again, we summarize all the situation in a table including the supports for the claims.

Table 3: Situations restricting the focus 2 (data user)

Restriction to focus	Situation	Support
Description of how the focus is restricted	Description of the situation in which the focus is restricted	Reference to a data source
Data users do not want to participate	Data quality does not match with the users' requirements. Data users cannot judge the data quality by seeing the full dataset before buying.	*"As there is no intermediary to check the quality, we need a mechanism to convey confidence in data quality to convince data users to participate"* – Interview 2
Data users want to participate	Data is useful i.e. there is perceived added value for a stakeholder resulting from acquiring the data.	"*As a data user, I need to be able to find sensor data that actually benefit my business by for example improving my production predictions*" – Interview 3
Data users do not want to participate	Data buyers have to pay for getting the data on the marketplace. Data is available elsewhere at a lower cost.	*"Scarcity is one of the main driver for attracting data users. Having the data available for free on other websites would harm our marketplace*" – Interview 4
Data users do not want to participate	Transaction is sensitive for data users to the competition. The competition can see the transaction.	"*Privacy is also important on the data user side, since it may give insights into their strategic plans"* – Interview 3

There are two situations that we believe are important to consider for the design as they represent requirements from respectively the user and the provider side, and as they are highly dependent on the context. In addition, they can be solved using blockchain technology applications that we have presented in the previous chapter namely the token curated data and the access-control. We will apply the step 2 and 3 of the method to these situations and therefore define the relevant components to include in the system. The two situations are:

- *Situation 2.1: Data quality does not match with the users' requirements. Data users cannot judge the data quality by seeing the full dataset before buying.*

- *Situation 1.1: The data is sensitive for the data owner to some businesses. The data is shared with the business the data is sensitive to.*

6.4.3 Quality perception as a proxy for actual quality: basic system functionalities

Before going further into the analysis of the *data quality* situation, we introduce and describe a crucial element as part of the basic functionalities: the "blind estimation" of the actual quality. We introduce this part outside of the method framework as it is not a context-aware part, but a core element upon which other parts of the design are based. We also did not introduce this part in Section 6.2 about basic requirements of the marketplace since information about the context was required to introduce the data quality, perceived quality, and required quality.

Data users have some requirements about the data quality, as one of the situations restricting the focus is that data needs to create value for the user. Below a specific quality level, the user cannot properly extract insights from data and therefore the user is not willing to participate. In practice, the data user does not have a specific quality level that can be quantified, it is more on a qualitative level (quality level can take values such as "low" or "high"). Nevertheless, we will use a quality level threshold as a requirement for the user. For instance, the data user may require a "high" level of quality, would therefore reject data showing a "low" level. This quality requirement could be estimated as a function of the factors aforementioned (number of data points, types of sensors, outliers identification and correction, etc…). Fox et al. (1994) provide a more extensive method for assessing data quality.

As a data user cannot judge the data quality by seeing the full dataset before buying it, they need to rely on a perception of the quality. So far, as we have not designed yet ways to convey an accurate perception, the quality perception is only based on metadata that complements the data. In other words, the perceived quality is defined by the description of the dataset provided by the sensor owner. This description could include samples of the dataset.

Based on this discussion, we can refine the situation into three parts:

Situation a
Data users have quality requirements and the quality of the data presented is lower than the quality requirement.

Situation b
The quality perception does not reflect accurately the quality level. In a system where the user knows that the quality perceived does not correspond to the actual quality, it does not matter whether the quality perceived is higher or lower than the quality requirement. The lack of accuracy is therefore a sufficient condition for not participating.

Situation c
Data users have quality requirements and the perceived quality of the data presented is lower than the quality requirement.

We can write these situations using predicates, with Q_1 being the quality level, Q_2 the quality perceived, and Q_3 the quality required:

Table 4: connections between the different types of data quality

Situation a:	Situation b:	Situation c:
$Q_1 < Q_3$	$Q_1 = Q_2$	$Q_2 < Q_3$

We observe that if $Q_1 = Q_2$ the predicate for situation c gives the same value as for situation a. This means that if we manage to change situation b such that the quality perceived is an accurate description of the real quality, then we are left with the situation c to solve. By "solving", we mean designing the necessary components to sense and adapt to the context. We now introduce the basic functionalities that ensure this accuracy and later we will be using only the perceived quality for the design of sensors and adaptors.

Once some data buyers have acquired the data, their judgment about the quality of the data should be considered and processed to have an accurate estimation. In addition to this reputation mechanism, there should be a way to estimate to what extent is the data provider confident in the quality of the data he suggests.

One measurement technique is using an external data auditor who receives the data by the provider and certifies the data quality. However, this means coming back to more centralized model by giving data access to a company that may have different incentives and act opportunistically. We have previously concluded that such a practice comes with important disadvantages such as a need for trust in this specific type of intermediary and should be avoided. Instead, we suggest using two types of measurements which together constitute an indicator for the quality:

1. *Reputation mechanism*

The first value of the quality factor is the assessment given by users. Users who have transacted with a data provider can give a rating to the provider, or to the dataset. The rating of the dataset could be the aggregate of several ratings for different parts which together

define quality. Some examples of rating questions: 1) Is the dataset provided with the same format as stipulated in the description? 2) Are the measurements accurate? 3) Are there as many data points as claimed? 4) Is the source of the data reliable? 5) Is the method used to replace outliers or missing data points rigorous? Most of the users do not have the knowledge nor resources to proceed to such an evaluation. However, some users can be expected to do so, such as users who collect data from different sources (and potentially produce themselves) to cross-validate results.

Marketplaces using reputation mechanisms face the "cold start problem" which says that it is difficult for new stakeholders to enter the marketplace as they do not have a reputation yet and therefore face a trust challenge. In addition, the decentralized data marketplace is a public network, meaning that everybody can access it. This can lead to false claims about the quality of a dataset, such as ratings that can be found about products on the internet. We therefore need a stronger mechanism to improve claims reliability. A possibility to reach this objective is incentivizing stakeholders to stake value in order to prove the data quality, via the use of tokens.

2. *Confidence in the dataset quality from the data provider*

The provider himself can be more or less sure about the quality of his results and should be able to not only mention this level of certainty in the description but there should be a more effective way to show this confidence, by involving some level of commitment. We suggest using the token curated data proposed in Section 5.4.3. As a short reminder, in a token curated data marketplace, data providers stake some tokens that are kept in escrow via a smart contract, meaning that they are blocked and cannot be accessed anymore. Other participants can bet for or against the dataset as well. As an output, each dataset is associated with 1 to 3 variables: the number of tokens in favor, the number of tokens against, and the difference between the two. Whether we should use 1, 2 or 3 variables depends on the *fungibility* of the tokens. Token fungibility refers to assets that can be interchanged with other assets of the same type (Entriken et al., 2017). In our case it depends on whether a token is worth another one. It could not be the case, for example if the data provider has a token from a different type, or stakeholders could be given more token power than others based on pre-established rules. For instance, actors that have successfully bet against datasets that had poor quality several times could be rewarded with more important tokens. They could also just be compensated with more tokens in a fungible tokenization system. In any case, the governance model of the platform must be transparent about these choices and should be voted upon by the stakeholders and/or defined clearly by the developers before the emergence of the marketplace, by making the code open-source. The data marketplace developers should not have prioritized access to tokens or to modifications of the token-related rules; a consensus is required according to the decentralized nature of the marketplace. In other words, the token economics should be immutable once established, but if a change is required they should be voted by the majority of token holders and not by the group of developers. Now that we have

proposed a solution to estimate the quality via a blockchain-based application, we initiate the analyze of the situations established in 6.2.2.

6.4.4 Analyzing data

Analyze for situation 2.1: The system proposes to the user a data set that does not meet the quality requirements of the user. The user is provided with metadata that conveys a perception by the user about the data quality.

For this first situation, the objects involved are the data, data user, data provider, the perceived quality level of the data, the required quality level of the user. The table below contains these objects as well as statements that can be made about these and the context elements which follow from the statements. The table will be the input for the next step, as the context elements listed will be sensed or will adapt. Following our discussion in 6.4.3, we assume the perceived quality to be an accurate proxy for the actual quality, and this statement is based on a basic functionality of our design. As a consequence, it is not part of the context to determine the accuracy between perceived and actual qualities. We will therefore work only with the perceived quality Q_2, but the reader should keep in mind that is also reflects accurately the actual quality Q_1.

The context element $proposesData(D, U, Q2, Q3)$ may require further explanations: It represents the fact that on the data marketplace, once the user queries some data by looking for keywords (e.g. types of sensors), some data should be returned and appear on the user interface.

Table 5: Analysis of the data quality situation

Situation 2.1: Data which quality does not satisfy requirements of user is proposed. The user is provided with metadata that conveys a perception about this data quality.		
Objects	Statements about the objects	Context elements
Data User (U) Data (D) Perceived quality level (Q_2) Required quality level (Q_3) Data Provider (P)	-U is a data user -P is a data provider -D is data (dataset) -Data provider uploads data - Q_2, Q_3 are levels of quality -Data has a perceived level of quality Q_2 -User requires a level of quality Q_3 for D -Data is proposed to the user	$isDataUser(U)$ $isDataProvider(P)$ $isData(D)$ $Uploads(P, D)$ $isQualityLevel(Q2)$ $isQualityLevel(Q3)$ $hasPerceivedQualityLevel(D, Q2)$ $hasRequiredQualityLevel(U, D, Q$ $proposesData(D, U, Q2, Q3)$

	-The perceived quality is below the quality required	$Q2 < Q3$

Analyze for situation 1.1: The data is sensitive for the data owner to some data user. The data is shared with this data user.

In this second case, the objects involved are the data, data provider and user, with a flow of data from P to B. We denoted the user with B for "Business" instead of U to emphasize that in this case the provider is not willing to share the dataset with the user. The flow is characterized by a sensitivity level. There is therefore an object relationship between these and a level of sensitivity. This leads to the definition of the table below, with several context elements. We can observe that the context element $isSensitiveTo$ takes several variables as inputs. This context element is a predicate taking the flow as an input, which includes some contextual information about the data transfer. The specific data flow is also part of the context since the data provider may not be willing to share the data in flow F_1 but is in F_2 e.g. if F_2 happens several months later, when some related projects have ended, and the data is not sensitive anymore.

By listing the objects and statements about the objects as we did for situation 2.1, we can define the context elements that will be used in the following parts.

Table 6: Analysis of the data sensitivity situation

Situation 1.1: The data is sensitive for the data owner to some data user. The data is shared with this data user.		
Objects	Statements about the objects	Context elements
Data provider (P) Data user (B) Data (D) Data flow (F)	- P is a data provider - B is a data user - D is a piece of data - F is a data flow -Data provider P uploads D -Business B downloads D - D is sensitive for P regarding a data flow F towards B	$isDataProvider(P)$ $isDataUser(B)$ $isData(D)$ $isDataFlow(F)$ $uploads(P,D)$ $downloads(B,D)$ $isSensitiveTo(D,P,B,F)$

6.5 Determining the components needed to sense and adapt to context

6.5.1 Determine what adaptors are needed

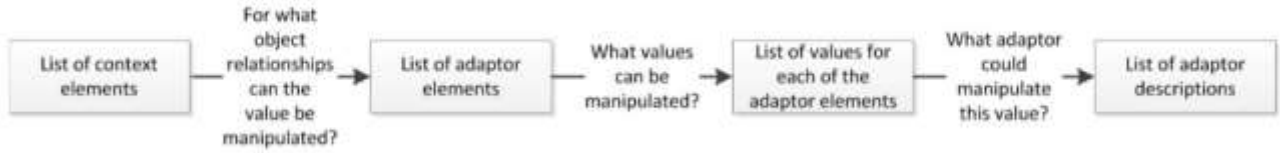

Figure 22: Determining what adaptors are needed (van Engelenburg et al., 2018)

Analyze for situation 2.1: Data which quality does not satisfy requirements of user is proposed. The user is provided with metadata that conveys a perception about this data quality.

We now need to identify which of the context elements should be manipulated, in the sense that the system should change predicates to match with the designer goal. The restriction of the focus to "data users do not want to participate" when data with insufficient quality are proposed indicates that there is a negative context relationship. In fact, as defined in the methodology (van Engelenburg et al., 2018), "A *negative context relationship* restricts the focus to a value that does not conform with the goal of the designer." In this case the designer wants more stakeholders from both sides to participate in the market to increase the value of the market. As it is a negative context relationship, we need to have at least one context element with a different value. In this case, we can change the value of $proposesData(D, U, Q2, Q3)$ such that this predicate becomes false when the quality is inferior. In other words, if the perceived quality is inferior to the requirement, the system should not recommend the dataset. The first adaptor is therefore the presentation of the dataset to the user.

Adapting element: $proposesData(D, U, Q2, Q3)$

To change the value of this element, we need to manipulate variables from the set {D, U, Q_2, Q_3} such that the predicate $proposesData(D, U, Q2, Q3)$ is false. Nevertheless, if the perceived quality is above the requirement, then the data should be proposed to the user.

We cannot change the user and his data requirements. We also cannot modify *directly* the perceived quality of the dataset by arbitrarily changing the metadata, as the system needs to be fair. The system is fair if the perceived quality cannot be manipulated by the system to steer the user towards buying the dataset. However, the system can manipulate *indirectly* the perceived quality, by changing the actual quality which will then reflect on the perceived quality (because of the accuracy assumption we previously made). There are therefore two remaining possibilities which are respectively manipulating D and Q_2 (through Q_1).

1) *Manipulate the dataset variable with a recommendation system*
   There should be a recommendation system that explores the databases by relevance of keywords and picks a dataset. However, if the quality is below the quality required the system should adapt and look for another dataset with a better quality. The recommendation system should therefore rank based not only on the relevance in terms of keywords, but also in terms of quality. The exact proportions for both factors could be determined experimentally. The adaptor component is therefore the recommendation system that adapts the ranking based on the quality, and the value manipulated is the suggested dataset D. As the component needs to connect directly or indirectly with all objects in the context element to manipulate it, it needs to connect with the dataset, with the user and her requirement, and with the data perceived quality. For the user, this is done by providing him with the dataset via an interface on the website or application. The adaptor connects with the perceived quality and the required quality by checking for all data if the perceived quality is above the quality required.

2) *Manipulate the quality of the dataset by outsourcing data curation*
   The data provider may not have the resources to improve the quality. However, there are data scientists in the ecosystem that should be able to provide curation services to the data. However, to not conflict with another situation that restricts negatively the focus, namely "the data owner loses ownership of data", we need the data provider to agree with the improvement. The mechanism could be:

   i. Data owner provides the raw sensor data D_1 to the marketplace.
   ii. Data scientist buys D_1.
   iii. Data scientist cures D_1 and thus creates a new dataset D_2, with Quality(D_2) > Quality(D_1).
   iv. When D_2 is sold, a smart contract automatically redistributes the revenues to the user and the data scientist, respectively the amounts corresponding to the raw data and the added value.

   This mechanism is a catalyst for co-creation of value, while conserving data ownership. As ownership is not only about profit opportunities but also about merit,

there needs to be a clear mention of who the raw data provider is. This is possible to achieve with blockchain since it keeps track of each transaction in the ledger that can be consulted by all parties. Therefore, the front-end (i.e. user interface) of the marketplace will automatically connect to the blockchain and retrieve the data provider pseudonymous to present it downstream in the data value chain, on the buyer side. If several data services are included, the same principle is applied in cascade, while keeping transparency since all operations are recorded. By doing so, the system also increases the overall transparency which was also one of the situations. In section 4.3.5 we explained that current data brokers exchange data among themselves, leading to a complex network where data traceability was a major challenge (cf. figure 15). With blockchain, as every asset transaction is recorded in a public ledger, we can add several data service layers and still maintain the transparency.

As mentioned above, the recommendation system will adapt based on the quality of data. However, it cannot access the data directly but only the metadata. We therefore need a mechanism that can sense the perceived data quality based on these metadata. Before introducing the sensors, we repeat the adaptor step to the situation 1.1.

Adaptors for situation 1.1: The data is sensitive for the data owner to some businesses. The data is shared with the business the data is sensitive to.

For the situation 1.1, the context relationship is also negative, since if the data provider uploads data that are sensitive but downloaded by a business on the "black list", then the focus is restricted to "data provider not willing to participate". We therefore need to have at least one context element that can change value. In this case we can restrict the access to data and therefore select the context element $downloads(B,D)$.

Adapting element: $downloads(B,D)$

We cannot change the dataset D itself because it has been uploaded by the provider on the data marketplace. However, the data provider can decide specifically which stakeholders can access the data. We can use the block-based permission system described in 5.4.1. As a reminder, this system allows the user to grant access by sending an access transaction to the blockchain with the public keys of user that can access the data. Data are stored encrypted on some databases connected to the blockchain via distributed hash tables but can then be queried by the data user after verifying that his digital signatures matches with his public key. The data provider can therefore decide on a very granular level who can access the data. The adaptor is therefore the access control system, which connects with the data provider through access transactions, with the data user as it contains his public key and can therefore verify digital signatures from him. It connects with the data via data transactions (upload and download).

6.5.2 Determine what sensors are needed

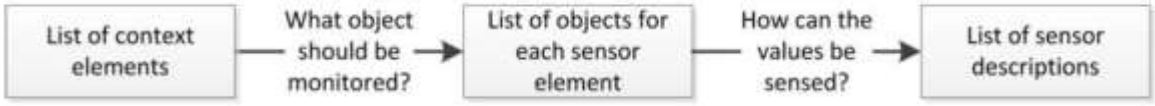

Figure 23: Determining what sensors are needed (van Engelenburg et al., 2018)

Situation 2.1: Data which quality does not satisfy requirements of user is proposed. The user is provided with metadata that conveys a perception about this data quality.

After having defined adaptor components for the focus "data users do not want to participate" and the situation 2.1, the following step is to look at other context elements and evaluate how their values can be sensed. As stated in the methodology (van Engelenburg et al., 2018), this part can be decomposed into four sub-parts:

1) Deciding which objects should be monitored
2) Finding a measurement for the values of the context elements
3) Defining which component (sensor) can measure
4) Evaluating the connections that the sensor has with the environment

The adaptor requires some type of feedback from the users to update the information about the data, but it has not yet been indicated how the quality perception of the users can be measured. Therefore, the object Perceived Quality level should be monitored, and its associated context element estimated: which must be sensed is $hasPerceivedQualityLevel(D, Q2)$.

$$\textit{Sensed element}: hasPerceivedQualityLevel(D, Q2)$$

We have previously defined the perceived quality as the number of tokens at stake on a dataset. This number corresponds to the measurement that must be taken, and the sensor could therefore be a token counter included in the smart contracts which keeps the tokens in escrow. The sensor will therefore periodically refresh the tokens associated to all data on the marketplace.

Finally, the fourth and last step is to evaluate the connections of the sensors with the environment. The sensor has to connect with all objects: the data, the data provider and users, the quality and information. The counter is connected to the data as it read its metadata (number of tokens). It connects to the users and provider via the interface by converting a number of tokens in a low to high metric.

We also need a second sensor to measure the quality required by the user about the dataset, since the system will need to verify $Q2 > Q3$.

Sensed element: $hasRequiredQualityLevel(U, D, Q3)$

If we want to compare Q_2 and Q_3, we need them to have the same metrics. So far, Q_2 is measured in number of tokens and Q_3 is a list of qualitative requirements about data properties (e.g. file format). We can proceed in two ways for the translation in another metric: (1) converting the data quality required in a number of tokens. (2) keep the required quality qualitative but detail, for each factor, if the dataset should rigorously meet the requirement or not. For instance, by indicating three level: not important, important, very important.

For the first option, if the data user is used to the platform, he will have an intuition of the number of tokens that he expects. However, it is more complicated than that: it is not only a function of the number of tokens, but also of the description. As a reminder, despite using several times a shortcut by saying that the number of tokens reflects the confidence in the quality, it actually represents the confidence in the fact that the dataset matches with the description given by the provider. A description showing a poor quality with many tokens staked still results in a poor quality. Therefore, it is better to not use this method.

Concerning the second option, the user details each feature requirement and gives an attribute on a scale ranging from not important to very important. The data provider should do the same: for each feature inserted in the description, a bet should be made. This enables to decompose the confidence in the quality in the specific confidence into respecting each of the feature mentioned in the description. The translation from number of tokens to confidence in the feature can be done using a benchmark i.e. looking at the average amount of tokens staked per feature on the whole network. This would evolve with time but remains fair since no one is deciding for others what the quality standards are. At this point, we need to be careful since it may become very complex the more features there are. There should be a limited number of possible features defined by the developers. For instance, it could be limited to the following features:

1. Format of the file
2. Date of acquisition
3. Outliers treatment (replacing, deleting, etc…)
4. Location of the sensor
5. Number of sensors and density

The data provider would write the corresponding value for each of these features in the description and splits his bet between these. On the user side, this implementation can be done by asking directly to the user conditions that need to be met (e.g. JSON file, > 1000 sensors with a density lower than 1/acre in real-time data). The component connects to the user via the user-interface, and to the data by setting an expectation about its quality.

The remaining objects also need to be monitored, namely the data user, the data provider, and the data itself. The user and provider values (i.e. who they are) is sensed when they log in the marketplace, while the data is a file sent to the database via the blockchain and therefore recorded via a data transaction.

Sensors for situation 1.1: The data is sensitive for the data owner to some businesses. The data is shared with the business the data is sensitive to.

In this case the main context elements that must be sensed is $isSensitiveTo(D, P, B, F)$. In this case we decide to monitor the object P, the data provider. Following the reasoning in van Engelenburg et al. (2018) applied to business to government information sharing, we identify that the data and the flow do not have an intrinsic sensitivity level, it depends on which business it would be share with and via which flow. It is therefore not helping to monitor the data nor the flow. We also do not monitor the business B since they have a conflict of interest and would therefore not claim themselves that sharing the data with them is sensitive. To measure the sensitivity, we can just ask the data provider P when the data are uploaded on the marketplace.

The second step is to provide measurements to know which data are sensitive, for which business and in which data flow. The data are already referred by their hash in the distributed hash table, the hash seems therefore a direct and relevant option as a measurement. Each business has a public key that can be used, and this public key is connected with each business name since they need to register and log in the marketplace. For a measure for the data flow, we use the data transaction ID. As with other public blockchain protocol, each transaction is assigned with an ID and stakeholders can find transactions based on this ID. However, a flow is just characterized by a data sender, an item sent (the data), a data receiver and a timestamp. All elements are already being measured except the timestamp. Measuring the flow can be reduced to measure when does the data exchange occur and we therefore use a time scale. Days for instance is a relevant time scale based upon the duration of business projects. This is also an element that should be asked to the provider as part of the sensitivity. As explained in the situation analysis, a data provider may be willing to share data with business B after a period of time has ended.

The sensor connects with the data by sending the sensitivity information to the blockchain as part of the metadata. It connects to the user by sending a sensitivity request when the provider uploads data. It connects with the business (user) by retrieving his information (name & public key) in the blockchain, and with the dataflow by checking the timestamp of the user data request, which is also retrieved from the blockchain as the user has broadcasted a data transaction to query data.

6.6 Determining the rules for reasoning with context information

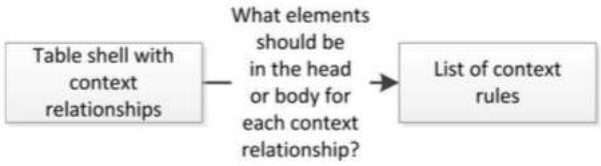

Figure 24: Establishing reasoning rules (van Engelenburg et al., 2018)

The input of the reasoning component are facts measured by the sensors and expressed by ground literals. These inputs are translated via the reasoning rule to another ground literal that indicates which value the adaptors should have in order to give to the focus the value that the designer targets.

Reasoning rules for situation 2.1: Data which quality does not satisfy requirements of user is proposed. The user is provided with metadata that conveys a perception about this data quality.

The sensor generates facts, which are a level of confidence in the quality of the tokens by the provider and potentially other stakeholders from the system. The final step is to reflect upon the reasoning component which is taking these facts as input and which outputs are also facts which are the value that adaptor elements.

As a reminder, in the two previous steps two components have been proposed, which can be summarized as:

- A *sensor* component which counts the number of tokens at stake, diminished by the number of tokens opened for claims against a dataset. This token counter is part of a smart contract where the tokens are kept in escrow, and its output value is written in the metadata (in the blockchain).
- An *adaptor* component which takes a fact as an input from the reasoning component and updates the dataset that is presented to the user, by modifying the user-interface. This dataset presented and its associated quality influence the focus i.e. their willingness to participate in a transaction.

The remaining part is about the rules for reasoning which take facts based on data generated by the sensors expressed as ground literals. The output is also represented as ground literals and are commands to the adaptor:

$$\neg proposesData(D,U,Q2,Q3) \leftarrow isData(D), isDataProvider(P), isDataUser(U), Uploads(P,D), hasPerceivedQualityLevel(D,Q2), hasRequiredQualityLevel(U,D,Q3), Q2 < Q3$$

As long as the perceived quality of the dataset is sensed (via the token counter) as being lower than the sensed (via asking to the user) required quality, the data should not be

proposed to the user. Now if the sensors see that a new data is uploaded, they should re-estimate the quality comparison, and if the perceived quality is satisfying, it should be proposed.

Reasoning rule for situation 1.1: The data is sensitive for the data owner to some businesses. The data is shared with the business the data is sensitive to.

As a reminder, in the two previous steps two components have been proposed, which can be summarized as:

- A *sensor* component which asks to the data provider who the data cannot be shared with, and for which period of time. Other facts measured are who the data user is, which dataset, is it uploaded on the marketplace, and when is the user willing to access the data (data flow).
- An *adaptor* component which takes a fact as an input from the reasoning component and restrict the access using a blockchain-based permission control system.

The remaining part is the rules for reasoning which takes facts based on data generated by the sensors expressed as ground literals. The output is also represented as ground literals and are commands to the adaptor:

$$\neg\, downloads(B,D) \leftarrow isDataProvider(P), is\ DataUser(B), isData(D), isDataFlow(F), uploads(P,D), isSensitiveTo(D,P,B,F)$$

6.7 Components integration

In this section, we combine the components previously described to form the marketplace system. And we illustrate the data exchange process through a scenario (i.e. an example) and using Business Process Management & Notation diagrams. The paragraphs below introduce the stakeholders in the scenario. Following sub-section describe the final system representation by illustrating the process of respectively uploading data i.e. supply side (Section 6.6.1) and downloading data i.e. demand side (Section 6.6.2). Each paragraph has a title that corresponds to the relevant part on the diagrams.

Alice works for a major car company. She is a sensor owner as each car is equipped with sensors measuring some technical characteristics when driving. These sensors could be a camera at the front, and an accelerator. As a result, they have datasets that show the current acceleration (or deceleration) based on pictures taken by the camera. Alice (and her company) is willing to monetize this data. However, she does not want to share it with direct competitors since they could try to find flaws in the cars that would hurt the company image e.g. they could realize that the brakes are not resistant enough as the deceleration

performances of cars decrease over time. They could also pinpoint weak resistance of the cars by observing pictures illustrating crashes (e.g. by looking at the plastic deformation).

Bob works for a startup working on software for self-driving cars. They are interested in feeding a machine learning algorithm that aims at determining the right acceleration/deceleration based on the environment (what is observed on the picture). They are therefore highly interested in the datasets collected by Alice. Bob wants to provide a very accurate software and therefore needs data collected every second and in high definition (e.g. minimum 1080x720 pixels). He is a data user. Bob is not seen as a competitor by Alice, since she believes her company can acquire or buy his software if it appears to show positive results.

6.7.1 Architecture for uploading data

Collecting sensors data & Defining sharing rules (including description)
Alice realizes that she possesses data that may be useful for other companies and decide to share these data using the decentralized platform (because she does not trust centralized ones).
Based on her preferences, she defines a set of rules though the user-interface of the system. The set of rules include: price, time availability (how long are the data available for?), the possible public keys of the actors that she is already granting access to, as well as possible complementary information such as a data contract. In this case, she may not know about Bob's startups and that they would be interested into her data. In this case, she just does not specify any public key. She also gives a description of the dataset e.g. date, data collection process, types and locations of sensors. In this case, she would write specifications like "Acceleration (m/s^2) of [car model] and associated driver vision illustrated by images (1080x720 pixels) taken with a [camera model], every second".

Uploading & storing data, updating hash table, including transaction in block and diffusing
The data and metadata are then uploaded to the blockchain. More accurately, the blockchain automatically stores the data (encrypted or not) in a (distributed) database. As the blockchain has the storage limitation previously mentioned, the data is not stored on the blockchain but is hashed and the hashes are stored in blocks. The two are connected via the distributed hash table introduced in the previous chapter. As there is a new piece of data stored, the blockchain updates the distributed hash table. This update is inserted as a transaction, together with the metadata of the data, in a block. The block is then created. Depending on the specifications of the blockchain, the blocks should have a certain size and therefore can include a certain amount of transactions. Once created, the block is added to the blockchain and this is spread to the network that agrees upon the new state of the network via the consensus algorithms (which also depends on the specifications of the blockchain protocol used), after verification of the digital signatures.

Staking & storing tokens in a smart contract
When the data are uploaded, the data provider (Alice) can (and should) stake some tokens in order to show his confidence in the quality of the dataset. Tokens are kept in escrow by the blockchain which governs a smart contract, and this is recorded via a new transaction appended to a new block. The new block is then diffused. We notice that Bob, our data user, does not participate to this part of the process, the BPMN has therefore only two lanes accounting for the data provider and the blockchain-based marketplace.

The data uploading process is represented on Figure 24.

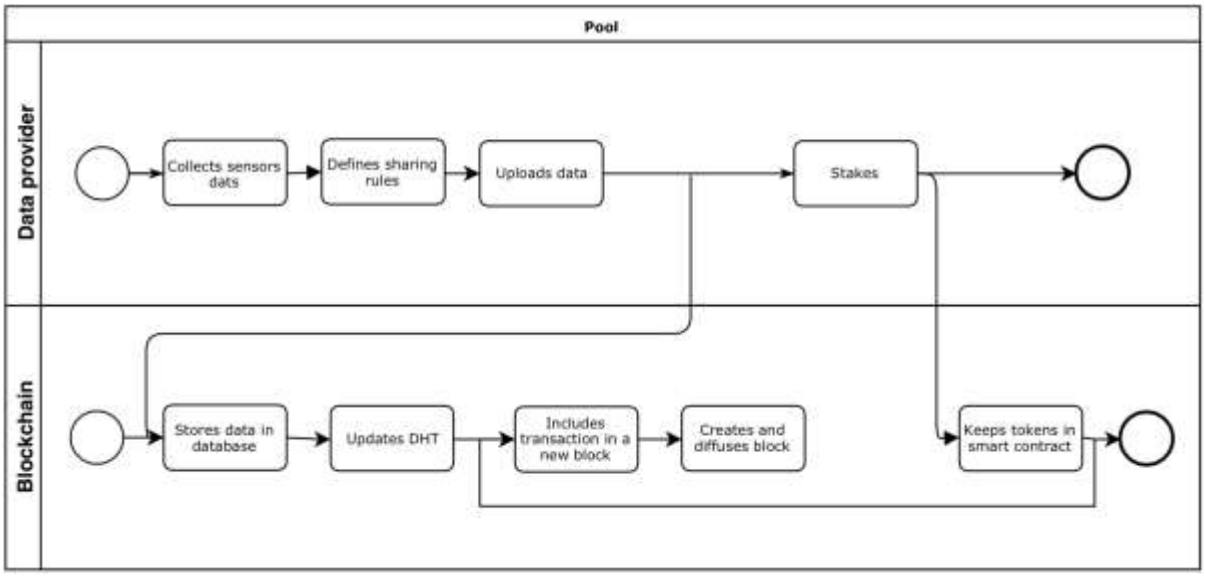

Figure 25: BPMN representing the data upload process on the marketplace

6.7.2 Architecture for downloading data

Data request
When the data user has defined his business needs and the necessary data and associated quality, he introduces a request through the platform. In this case, Bob realizes that his software will need to control the speed based on the events happening outside of the car. Therefore, Bob thinks about developing a machine learning algorithm taking as input these elements and the associated acceleration. He needs a high number of samples to train this algorithm. As an output, the algorithm will give a model that analyzes the information in the pictures and translates into the right acceleration (e.g. using neural networks). Bob enters some keywords in the data search bar, on the user application. For example, he could try the combination "car + acceleration + camera".

Browsing data catalogue, checking quality satisfaction

The front-end of the platform is connected to the blockchain that receives this request and starts browsing headers for the closest metadata using keywords. Based on the expected quality and based on the quality of the stored data which is sensed by the token counters, the blockchain selects or not the data. This is an iterative process since when a dataset does not meet the quality requirement, the recommendation system needs to look for another one until this is met. If no satisfying data can be found, an error message is returned. When a dataset with sufficient quality is detected, the blockchain checks the access. Note that there is no reason for checking the quality before the authorization. The best choice minimizes the overall process duration, which depends on the protocol implementation.

Access permission, recommend data

The access control is done by checking which public keys have been accepted by the data provider, based on the decentralized permission component as described previously. If the corresponding public key of the data user appears, then he is granted access to the dataset. If not, a request is sent to the data provider (the notifications can be disabled by the data provider for user-experience purposes). Alice receives this request and analyzes Bob's information. If, based on the business identity contextual element and the sensitivity of the data, she judges that the data can be shared with Bob, then she grants access. Otherwise, the recommendation system must look for another dataset and goes back in the loop.

Downloading, verifying quality, staking, storing tokens and updating rankings, and using data

When access is granted, the user can download the data. At this point he is therefore able to see if the data corresponds to what was stated in the description i.e. if the data has the quality that matches with the perceived quality. If it is the case, then the data user will use the sensor data for his business operations. Bob will be able to feed and train his algorithm. However, if it is not the case, then the user will stake tokens against the dataset to show his disapproval of the quality. At this point he also stakes tokens (to avoid Sybil attacks and false claims i.e. to avoid the fact that many participants just vote randomly against datasets) that counterbalance Alice's. The blockchain stores the tokens in the smart contract and decreases the counter, which automatically leads to adjusting the quality perceived and ranking provided by the recommendation system.

The data downloading process is represented on Figure 25.

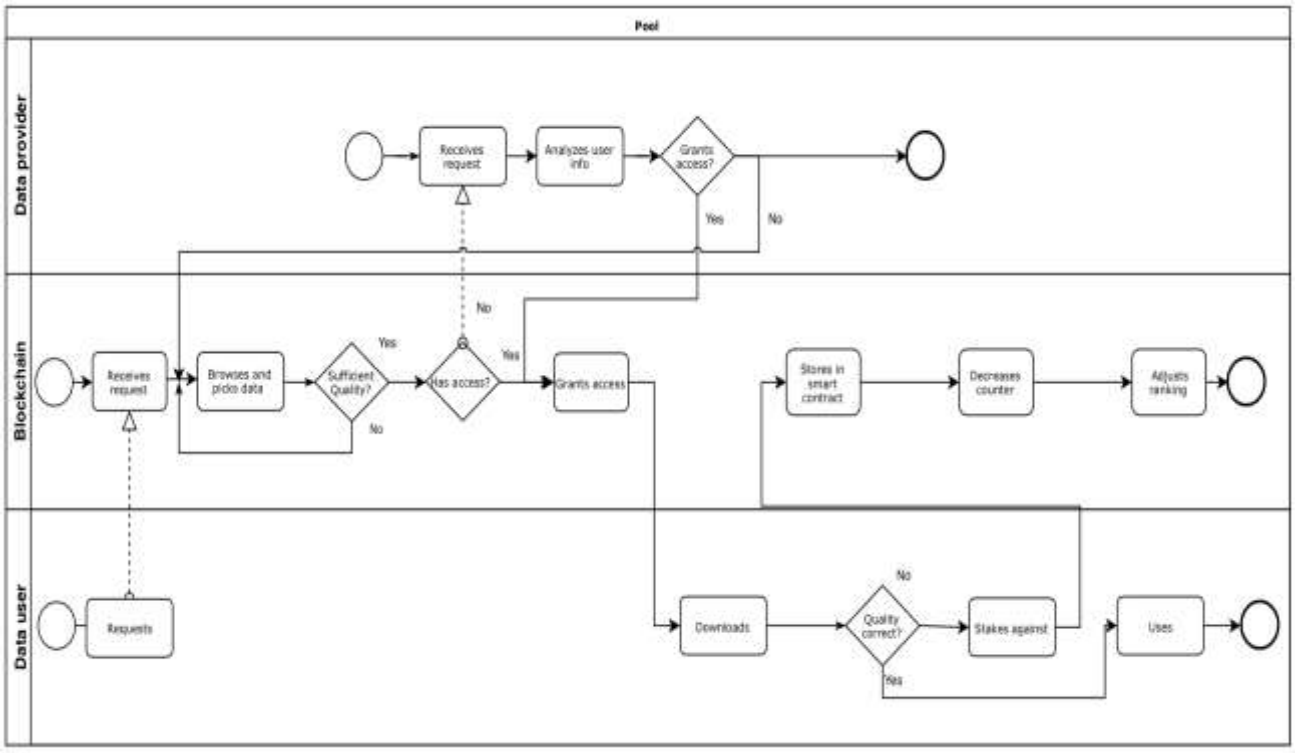

Figure 26: BPMN of the data user request and download processes

6.8 Design assessment

In this last section, we evaluate our decentralized data marketplace by discussing first the application of the method, and then limitations of the system including the missing elements. We finish this chapter by mentioning the strengths of this model and how it contributes to the main goal of reducing central firms' control while respecting stakeholders' values.

6.8.1 Methodology discussion

The data sensitivity-related situation has been selected as a negative relationship limiting the focus to the data provider *is not willing to participate*. However, after having applied the method and assessed it, we realized that based on the framing of the expert it would have been more accurate to use a positive relationship and restrict the focus to *is willing to participate*. Framing "*Privacy, security and control of data increase the willingness [of data providers] to share their data."* - Roman & Stefano (2016). Nevertheless, it does not have a significant impact on the system, since the adaptors and sensors are the same, only the reasoning rule differs.

The second point to emphasize is that for the design phase we have made the assumption that the perceived quality was an accurate representation of the actual quality. The argument used to justify this assumption was the choice of token curated data since it was proposed by several researchers in the field. However, none has implemented and successfully tested this

mechanism in real settings yet. There is also no behavioral research confirming the validity of this technique e.g. no game theory considerations. For such a complex process involving so many stakeholders with different values, we believe that testing is important to guarantee internal validity.

A third point that could have been improved is increasing the number of experts interviewed. Four interviews are a relatively small sample. In addition, the first interview was taking at the very beginning of the thesis and was therefore too unstructured. As a result, the discussions were not always relevant with the focus of this research. Finally, the two last interviews could not be recorded for technical reasons. Despite the extensive notes taken, it may have led to a loss of information.

6.8.2 Missing elements

The systems proposed in 6.5.1 and 6.5.2 are not sufficient to build a complete decentralized data marketplace, since some important elements are missing. Here are the main elements and paths suggestions for future research.

Data Storage

We have defined that the data should not be stored in the blockchain as it does not have the necessary storage capacities, and we have therefore connected the blockchain with "a (distributed) database." The ambiguity was left there since we did not investigate the different storage possibilities, their advantage and inconvenient beyond the fact that we need to protect data from firms owning these data storages. Typically, cloud providers offer easy to deploy storage possibilities to data providers, but again this implies more control by the central company (Kshetri, 2013). Therefore, we suggested not using a privately-owned cloud, but it is also valid for databases from third-parties. However, if the data is encrypted (and replicated to protect against the deleting attack), they can be stored on such solutions. We recommend to research further about storage solutions to be included in our system, and more specifically comparing distributed storage systems, clouds and traditional databases.

Conflict resolution

The token curated model is based on the assumption that there is an effective mechanism to resolve conflicts when some actors stake against and other for a dataset. The conflict can be resolved publicly (but that would entail making the data available), via some representatives or via an oracle. These different solutions offer a dilemma between effectiveness/feasibility and the presence of intermediary/authority. More research should be done about the optimal balance between the two extremes in the context of a data marketplace.

Data reproduction consequences and protection

We have also not discussed what happens when the data buyer side downloads data and replicates it illegally. This is a problem that can be studied from different perspectives e.g. by the use of technology or by the use of law.

Concerning protection with the technology, it is clear that from the moment the user can see the data, it is possible to copy it (although it may be cumbersome). In Section 5.4.4 we have suggested using homomorphic encryption to train models on data without making the data available to the model developers. However, more research is required about how to apply this method in this context, and what would be the limitations and potential attacks.

Concerning legal protection, law researchers are invited to reflect upon which contractual arrangements should be made to protect the data user. We suggest building on the idea that as all data transactions and access control transactions are recorded in the blockchain and as the blockchain is immutable, the data providers will always have a clear proof that they have not granted access to a party by consulting the blockchain. Therefore, if a party is found using the data and a claim is opened against him by the data provider, legal action can directly be taken. Following the same logic, if an actor has granted access to a data user and tries to open a claim against him arguing that the data was not shared with the user via the marketplace, the data user can protect himself by pointing at the transaction in the blockchain. The blockchain therefore guarantees a consensus about the truth and forces the actors to behave correctly, which indirectly creates trust in the system.

In addition to the legal responsibility, it is important to discuss the moral responsibilities of the data marketplaces stakeholders. As it is a more decentralized system involving peers almost directly, data users and providers have a greater control and may be confused about what are their responsibilities and what are responsibilities from other stakeholders such as the platform developers. For instance, who is morally responsible if the network is hacked? If a user handles wrongly an unclear smart contract provided by the developers? We believe that as with most other blockchain-based and non-blockchain-based decentralized systems, liability and moral responsibility cannot be assigned to a central party directing the process anymore, or at least not to the same extent as centralized systems. Therefore, there should be more research about defining frameworks for attributing responsibilities in these settings, starting with responsibilities for the decentralized data marketplace.

Economics and pricing mechanism

As a marketplace, there are a number of economic considerations that need to be taken, such as how to ensure liquidity? Should pricing be dynamic (e.g. adjusting to offers on other marketplaces)? The marketplace providers also need to get some rewards for building the marketplace and coordinating a decentralized governance. It is important to propose and compare several business models, and see which one to select based on a set of criteria. A typical model that is taken in e-commerce platform is to take a fee for each transaction, generally by charging the user side.

Real-time data

In the previous sections, we described the process for exchanging datasets. However, it is also possible to share data in real-time, by connecting to data streams. In this case, the data are not stored in the database, but the blockchain connects with API to retrieve and transfer data as fast as they are updated (plus latency time). The scenario below gives an example of such real-time exchange.

Alice is a sensor owner working full-time for a car-sharing company. Her mobile constantly tracks her position i.e. in real-time. She wants to monetize this information using the data marketplace. As she is aware (or at least she considers possible) that her driving behavior is not fully safe, she does not want her insurance company to get this data. She is afraid of individual insurance pricing that would translate in more expenses for her.

Bob works for a real-time route optimization company, which computes the fastest route by aggregating the positions of drivers such as Alice and evaluating the traffic density. Bob's company offers a premium service, in the sense that it is more expensive than competitors' offerings, but it is much more accurate. As a consequence, the data they require must include datapoints sufficiently close to each other on the time dimension e.g. every second.

We have not discussed how to implement the connection with the API. Additional research focusing on decentralized data exchange could focus on this possibility, and the potential extra challenges that would arise.

6.8.3 Strengths of the model

Despite the missing elements mentioned above, our proposition is a base for building a complete decentralized data marketplace. It respects the initial goal of having no data manipulation or control by a central firm.

It also respects the values from both types of stakeholders that based on our interviews and literature review were crucial: data providers can *monetize* their data, while keeping *ownership*. In fact, only with the private can someone change the ownership of data as it is necessary to show a valid digital signature. Therefore, the only possibility for one to lose ownership is to share one's private key. As with most blockchain applications, the blockchain design itself is secure, however there are risks of private keys theft when using complementary services. As an example, some applications may provide private keys management for users. The users should make sure they trust these application providers or should not use third-party applications. This is the same principle as with Bitcoin: it has never been hacked, but some third-party wallets and exchanges where bitcoin can be stored were. There is also *downstream transparency*, as each transaction is recorded in the blockchain. Therefore, the data provider knows when someone wants to access their data. As the latter had to identify himself, the identity is known to the data provider.

On the user side, there is *upstream transparency* since the user knows who posted the data as it is recorded in the ledger that can be consulted openly. Data *quality* is improved thanks to the staking mechanism. Finally, there is transactional *privacy*, since the buyer does not have to show publicly his identity when acquiring the data. Nevertheless, he should give his identity to the data provider as justified above with the downstream transparency value.

Chapter 7: Conclusions

As a reminder, the main research question of this thesis was:

"How can we improve efficient data sharing by reducing risks of opportunistic behaviors in firm-controlled data exchange mechanisms?"

We have answered this question by suggesting a blockchain-based, context-aware, and decentralized data marketplace architecture. This architecture removes the control of the firm over the data. Data are stored in databases and the blockchain includes pointers to localize these data, via distributed hash tables. The pointers are associated with metadata that contain a description of the dataset that is also stored in the blocks. Users can browse these descriptions and find the datasets they are looking for effectively thanks to a recommendation system. In addition to looking for keywords, this recommendation system uses a token mechanism to verify the data quality, without having to give data access to the firm controlling the marketplace or to the user. This token curated quality mechanisms incentivizes data providers to stake tokens on their dataset to show their confidence in the quality. The tokens are kept in escrow in a smart contract as long as the data are not proven to be invalid, in which case they are given to stakeholders who opened a claim against the data quality. To open a claim, stakeholders need to also stake tokens in order to protect the system from false claims and Sybil attacks. The data can be accessed by the user thanks to a decentralized permission system. This system checks the public keys of data users requesting the data and see if there is a match with the public keys authorized by the data provider. Data users can request access to data providers, who have a mobile application to quickly get notified, and accept or reject the request. As the developers of the marketplace do not have access neither do they host data anymore, opportunistic behaviors that could result from conflicting commercial interests are reduced. Data sharing is therefore improved in the sense suggested in the research question. We finally note that it is an *efficient* process since it is a marketplace and therefore it offers many-to-many data exchange possibilities, contrasting with one-to-one data sharing between two entities such as sending data via email.

We have articulated this research question in several sub-questions. We first had to identify and understand the current salient features of data exchange mechanisms. We formulated and answered the two first sub-questions in Chapter 4.

Sub-question 1: What are current solutions used for data sharing between sensor owners and data users?

We answered this question by making the difference between individual exchanges (one-to-one) involving two parties via a data contract agreement, and scalable 'solutions' (many-to-many). As the main research question emphasizes the efficiency, we modified the scope to

only include the latter type of solutions. Within this category, we have observed three main ways of sharing data in the literature: data brokers, privately-owned and open data marketplaces. A common characteristic that we could find in all these solutions is the fact that there is always a few (generally one) companies owning and controlling the data exchange. From the interviews that we had and based on the literature review, we concluded that it was leading to a trust problem. This was the topic investigated by question two.

Sub-question 2: Why is the trust in a company controlling the data exchange a problem?

The trust problem comes from the commercial interest that the firm has, leading to centralized data storage, lack of transparency about the data value chain, both downstream (on the user side) and upstream (on the provider side), loss of data ownership and control, and no guarantee of fair pricing. This led us to realize that there was a significant risk because of this misalignment of objectives between the central firm controlling the marketplace and its users. This could lead the firm to share data with third parties or use it for its own interest, which impacts negatively the willingness to participate of all participants. As a consequence, data sharing is limited.

To resolve the centralization problem, we looked at blockchain technology since it has been used in other fields to increase decentralization. We reviewed the literature to identify the relevant blockchain parts that should be used, and how.

A context-aware system is required to deal with the complexity of a decentralized environments, where the system requires more automation and therefore needs to be able to sense and adapt to contextual elements. To build the context-aware system, we used the design method proposed by van Engelenburg et al. (2018) which required us to understand what is part of the context.

Sub-question 3: Which parts of the environment belong to context?

We then entered the design phase and suggested blockchain components, before applying the design method to answer the main research question.

Sub-question 4: Which blockchain applications or properties can be used to achieve more decentralization efficiently in a sensor data marketplace?

The relevant blockchain applications that we selected are: smart contracts, decentralized permission control with off-chain data storage connected to the blockchain via distributed hash table, smart contracts, and token curated data markets. These applications are added to some basic blockchain functionalities like immutability, transparency and distributed aspect. In addition, we proposed homomorphic encryption as an enabling technology but were not able to integrate it. It was discussed and suggested for further research in Section 6.8.2.

With mainly the interviews, but also extensively by reviewing the literature, we answered the main question by defining two main foci. These two foci are crucial elements of a marketplace as they relate to the willingness of both supply and demand sides of the marketplace. 10 situations restrict these foci (respectively six for the data providers and four for the data users) and were summarized in tables 2 and 3. By applying the complete method to two of these 10 situations, several components (adaptors, sensors, and reasoning rules) have been proposed and integrated into a global system. The system has been described using Business Process Management & Notations diagrams which are the main outcome of this thesis and answers to the research question. As mentioned extensively above, risks are reduced because the owner of the marketplace has less control over the process, and users and providers are willing to participate which improves data sharing. The improvement is effective since it is a marketplace and therefore there are many-to-many data exchanges.

We finally evaluated our decentralized data marketplace and therefore answered sub-question 5 in Section 6.8. Missing elements were highlighted, namely the lack of research about which type of storage should be selected (databases, cloud, distributed databases?), the conflict resolution when parties disagree about the data quality (vote? Call to an external auditor?). We also highlighted that we did not consider how to deal with cases where data are illegally reproduced and how security could be improved even more using state-of-the-art cryptographic techniques such as homomorphic encryption. Finally, we did not discuss economics and pricing mechanisms of the marketplace, and the possibility to include data streams in real-time in addition to static datasets.

In the evaluation part, we also discussed some inaccuracies in the way the method was applied but resulting in a very limited impact since the physical components designed are the same. Only the reasoning rules differ. We finally ended the assessment by concluding that the model was meeting the goal of removing central firms (and their associated risks), and it met the values required by stakeholders: transparency, monetization, data ownership, privacy, and data quality.

Chapter 8: Discussion and future research

8.1 Generalization: suggestions for future research

8.1.1 Introduction

The decentralized data marketplace parts that we have proposed have been specifically designed for proprietary but non-personal sensor data, for data flows taking place from sensor owners to businesses. There is so far no scientific proof that the model can be used for other purposes, such as a personal data marketplace for the healthcare industry. In this section, we discuss several paths for expanding the model and proposes research orientations.

To verify and improve the external validity of the design, we recommend using analytical generalization (Gibbert et al., 2008). More specifically, the cross-case analysis method (Eisenhardt, 1989) proposes trying the design with 4 to 10 other cases, to evaluate to what extent our model fits with different phenomena and what should be changed.

8.1.2 Personal Data and GDPR

The design may be challenged when using personal data as it is subject to stronger regulations and personal risks. Michèle Finck has written about relations between blockchain-based systems and General Data Protection Regulation (Finck, 2017a) and between blockchain and the overall current regulations. (Finck, 2017b). We suggest using these papers as a starting point for modifying our system to support personal data.

8.1.3 Including non-human agents

One of the main promises of the internet-of-things is the faculty of communication between objects without human intervention. We could therefore extend the decentralized data marketplace for internet-of-things markets, where autonomous agents would buy directly data from each other through the marketplace, without human intervention. Hammi et al. (2018) have introduced the concept of *bubble of trust*, blockchain-based secure virtual zones where the data exchanges between devices is secure. Their exchange mechanism could be combined with a decentralized data marketplace to increase the matching possibilities between device providing and using data. Future research focusing on such marketplace should consider using their device authentication solution as its security has been tested in their paper.

8.2 Reflections

8.2.1 On the role of blockchain and tokens

In the complex and global world where businesses operate, it has become crucial to enforce trust between parties trading together. Blockchains and other distributed ledger technologies appears to be effective technologies to reach this objective, since every single transaction is recorded, leading to all network participants to agree on the state of the network i.e. the state of all trades. I trust blockchain; or rather, I believe that blockchain can enforce stakeholders to behave in certain ways, and therefore that blockchain is the catalyst for stronger trust relationships.

In addition, I realize the need for decentralization. However, decentralization has a major drawback: there is a lack of coordination between stakeholders. In the data marketplace, an illustration of this is the fact that there is no actor to "coordinate" actors to know which datasets are of a good quality. Tokenization enables to circumvent the coordination problem, as it does in the data marketplace when it represents the data quality. There is a high potential in blockchain-based tokenized ecosystem, which are also important to build decentralized autonomous organizations. These are agglomerates of actors working together, with the absence of a superior entity and hierarchy to run the organization. This is only possible if there is a careful design of incentives such that actors behave as expected. Leading blockchain thinkers such as McConaghy (2017) believe that we will see the emergence of a new discipline - "token engineering" or "incentive engineering" - focusing on the development of these tokenized ecosystems.

8.2.2 On the adoption of decentralized data marketplaces

During my thesis, I had the opportunity to discuss with many businesses and explain the concept of decentralized data marketplaces. Their positive reactions led me to think that, assuming the implementation matches with the theoretical promises in terms of user-friendliness, security, and speed, then decentralized data marketplaces will be globally adopted in the future. However, the technology is far from ready, with major problems such as blockchain scalability.

I could observe the first decentralized data sharing platforms being developed by several startups around the world, but as far as I know none has already a fully-working product that sensor owners and businesses can use. Following discussions with the developers of these marketplaces and based upon my own knowledge of software development and adoption, I see these platforms appearing on the market by the end of 2019 and crossing the chasm only around 2025. This is based on the fact that companies are still keeping data siloed, and not using current data exchange mechanisms despite the risks associated. It is unlikely that they do not use data marketplaces only because of the central firm, since there are cases in which going through a centralized data marketplace is acceptable (e.g. if a company trust a marketplace provider). It also shows a lack of awareness about the possibilities, and about the benefits of using these.

Blockchain is a very new technology even in scientific research as proven in Section 3.2.1. Understanding how and why a blockchain-based decentralized data marketplace differs from other marketplaces may require some time; and without understanding businesses may not be willing to participate.

On the philosophy level, the diffusion time estimation is also based on the adoption of decentralization. Decentralization has been seen as an objective by some for years already, with the rise of the internet and associated sharing platforms (e.g. Airbnb, Uber), the cloud, and now with blockchains. Nevertheless, it remains a shift in paradigm that society seems to be reluctant to embrace, as illustrated by the adoption problems that cryptocurrencies face (including Bitcoin). We are still far from the massive adoption of decentralized autonomous organizations to replace current companies and organizations.

Now, assuming that decentralized data marketplaces penetrate the market and reach significant market shares compared with other data sharing solutions, it remains to be seen which solutions will be adopted, and if several will co-exist. On the one hand, decentralized data marketplaces are subject to network effects (the more participants, the more attractive), which could give a significant competitive advantage to first movers. On the other hand, there is a clear tradeoff to make between decentralization, security, and scalability, as with other blockchain-based applications (Buterin, 2016). Developers targeting the "right" (as defined by providers' and users' preferences, which they may not be aware of currently, but these preferences may materialize later on) balance between these could be the successful ones.

Finally, it is impossible to forecast *who* will develop successful platforms. By "who", we specially refer to two possibilities: new players and/or incumbents. New players include startups, government and businesses that were not active in the data marketplace development before. They enter the market directly with a decentralized approach because they believe it is the right opportunity. Incumbents refer to the current data marketplaces developers e.g. Microsoft. These may perceive the threats that these decentralized data marketplaces represent and decide to initiate the move themselves. The Ocean Protocol Foundation is a mix of both, result of a merger between a new player (BigchainDB) and a centralized data exchange (DEX).

8.2.3 On the use of the context-aware method

The method used for this design-oriented thesis has helped to reach the design objective. The main strength of this method is its structure. It really forces the designer to carefully approach the problem in a structured way. It first requires articulating the main points to reach the goal i.e. articulating the foci. This helps steering the interviews to discuss what influences these foci. I observed that the interviews I took with the foci in mind were much more concrete than the others, in the sense that interviewees were really providing me with detailed reasons for participating or not in the marketplace. Then, it asks the designer to formalize the answer of interviewees into situation and context elements. Decomposing the design process in these

small pieces really help understand why certain design choices should be made. The clear illustrations using two running examples are another strength of this paper.

However, the method does not include everything that is required for the design of the whole context-aware marketplace. Mainly, the components are the outputs of the method, but there is little information about how to integrate the physical parts that are derived from these components. In other words, we connect elements within a set of sensors, adaptors, and reasoning rules. Based on this we target which physical tool (e.g. a screen to show data) should be part of the design. However, we do not know how to connect this with physical tools defined from other sets (e.g. with the permission system). I must admit that it is complex to include this in any method since these physical tools are very specific and it is hard to imagine which general guidelines could help integrating such specific artifacts. Indications about how to evaluate the design could also be inserted in the paper.

Finally, the vocabulary used in the method was quite complex to understand in the first place, and how does it relate with logic. However, the authors have changed some of the vocabulary which makes it easier to understand.

As a conclusion, I would recommend using this method to designers seeking to build context-aware systems.

References

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender systems. In Recommender systems handbook (pp. 217-253). Springer, Boston, MA.

Antonopoulos, A. M. (2014). Mastering Bitcoin: unlocking digital cryptocurrencies. " O'Reilly Media, Inc.".

Balazinska, M., Howe, B., & Suciu, D. (2011). Data markets in the cloud: An opportunity for the database community. Proc. of the VLDB Endowment, 4(12), 1482-1485.

Baldauf, M., Dustdar, S., & Rosenberg, F. (2007). A survey on context-aware systems. International Journal of Ad Hoc and Ubiquitous Computing, 2(4), 263-277.

Banerjee, S., Bolze, J. M. McNamara, & K. T. O'Reilly. (2011) How big data can fuel bigger growth. Accenture Outlook.

Banko, M., & Brill, E. (2001, July). Scaling to very very large corpora for natural language disambiguation. In Proceedings of the 39th annual meeting on association for computational linguistics (pp. 26-33). Association for Computational Linguistics.

Benet, J. (2014). IPFS-content addressed, versioned, P2P file system. arXiv preprint arXiv:1407.3561.

Bennati, S., & Pournaras, E. (2017). Privacy-enhancing aggregation of internet of things data via sensors grouping. arXiv preprint arXiv:1702.08817.

Bernstein, D. J. (2009). Introduction to post-quantum cryptography. In Post-quantum cryptography (pp. 1-14). Springer, Berlin, Heidelberg.

Bisgaard, J. J., Heise, M., & Steffensen, C. (2004). How is Context and Context-awareness defined and Applied? A survey of Context-awareness. Aalborg university.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. Science Fiction and Philosophy: From Time Travel to Superintelligence, 277-284.

Bulkowski, B. J., & Srinivasan, V. (2018). U.S. Patent Application No. 15/488,511

Buterin, V. (2014). A next-generation smart contract and decentralized application platform. white paper.

Buterin, V. (2014). On blockchains sharding. Retrieved from https://github.com/ethereum/wiki/wiki/Sharding-FAQs

Buterin, V., Wilkinson, S., Boshevski, T., Brandoff, J., (2014). Storj a peer-to-peer cloud storage network.

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian, 17.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: from big data to big impact. MIS quarterly, 1165-1188.

Chen, J., & Xue, Y. (2017). Bootstrapping a blockchain based ecosystem for big data exchange. In Big Data (BigData Congress), 2017 IEEE International Congress on (pp. 460-463). IEEE.

Chesbrough, H. W., (2006). Open innovation: The new imperative for creating and profiting from technology. Harvard Business Press.

Cohen, D., Crabtree B. "Qualitative Research Guidelines Project." July 2006.

Douceur, J. R. (2002, March). The sybil attack. In International workshop on peer-to-peer systems (pp. 251-260). Springer, Berlin, Heidelberg.

Eisenhardt, K. M. (1989). Building theories from case study research. Academy of management review, 14(4), 532-550.

Eisenmann, T., G. , Parker, G. , and Van Alstyne, M.W. , "Strategies for two-sided markets", Harvard Business Review 84.10, 2006, pp. 92.

Entriken, W,. ERC-721 Non-Fungible Token Standard. 2017. url: https://github.com/ethereum/EIPs/blob/master/EIPS/eip-721.md.

European Commission. "Communication from the Commission to the European Parliment, the council, the european economic and social committe and the committe of the regions – Open data: an engine for innovation growth and transparent governance"

Federal Trade Comission report (2013) What Information Do Data Brokers Have On Consumers, And How Do They Use It; committee on commerce, science & transportation US Senate

Federal Trade Commission. (2014). Data brokers: A call for transparency and accountability. Published: May.

Finck, M. (2017a). Blockchains and Data Protection in the European Union.

Finck, M. (2017b). Blockchain Regulation.

Fischer, G. (2012). Context-aware systems: the'right'information, at the'right'time, in the'right'place, in the'right'way, to the'right'person. In Proceedings of the international working conference on advanced visual interfaces (pp. 287-294). ACM.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In Early engagement and new technologies: Opening up the laboratory (pp. 55-95). Springer, Dordrecht.

Fox, C., Levitin, A., & Redman, T. (1994). The notion of data and its quality dimensions. Information processing & management, 30(1), 9-19

Gelfand, A. (2012). "Privacy and Biomedical Research: Building a Trust Infrastructure—An Exploration of Data-Driven and Process-Driven Approaches to Data Privacy," Biomedical Computation Review, Winter, pp. 23-28

Gentry, C., & Boneh, D. (2009). A fully homomorphic encryption scheme (Vol. 20, No. 09). Stanford: Stanford University.

Gibbert, M., Ruigrok, W., & Wicki, B. (2008). What passes as a rigorous case study?. Strategic management journal, 29(13), 1465-1474.

Gopalkrishnan V., Steier D., Lewis H., Guszcza J., and Lucker J..(2013) Big Data 2.0: New business strategies from big data. Deloitte Review. pp. 54–69.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), 8-12.

Hammi, M. T., Hammi, B., Bellot, P., & Serhrouchni, A. (2018). Bubbles of Trust: a decentralized Blockchain-based authentication system for IoT. Computers & Security.

Hart, P., & Saunders, C. (1997). Power and trust: Critical factors in the adoption and use of electronic data interchange. Organization science, 8(1), 23-42.

Haucap, J., & Heimeshoff, U. (2014). Google, Facebook, Amazon, eBay: Is the Internet driving competition or market monopolization? International Economics and Economic Policy, 11(1-2), 49-61.

Henderson, J. M., & Quandt, R. E. (1958). Microeconomic theory: a mathematical approach (No. HB171 H48).

Hoffman, D.L., Novak, T.P., and Peralta, M. Building consumer trust online. Commun. ACM 42, 4 (Apr. 1999), 80–85.

Jahansoozi, J. (2006). Organization-stakeholder relationships: Exploring trust and transparency. Journal of management development, 25(10), 942-955.

Janssen M., Charalabidis Y., Zuiderwijk A., "Benefits, adoption barriers and myths of open data and open government", Information Systems Management 29.4, 2012, pp. 258-268.

Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transportation Research Part A: Policy and Practice, 94, 182-193.

Karafiloski, E., & Mishev, A. (2017). Blockchain solutions for big data challenges: A literature review. In Smart Technologies, IEEE EUROCON 2017-17th International Conference on (pp. 763-768). IEEE.

Khine, P. P., & Wang, Z. S. (2018). Data lake: a new ideology in big data era. In ITM Web of Conferences (Vol. 17, p. 03025). EDP Sciences.

Koutroumpis P. Leiponen, A. E., "Understanding the value of (big) data," in 2013 IEEE International Conference on Big Data, 2013, pp. 38–42.

Koutroumpis, P., Leiponen, A., E & Thomas, L. D. (2017). The (Unfulfilled) Potential of Data Marketplaces (No. 53). The Research Institute of the Finnish Economy.

Kshetri, N. (2013). Privacy and security issues in cloud computing: The role of institutions and institutional evolution. Telecommunications Policy, 37(4-5), 372-386.

Lai, R., & Chuen, D. L. K. (2017). Blockchain–From Public to Private. In Handbook of Blockchain, Digital Finance, and Inclusion, Volume 2 (pp. 145-177).

Li, C., & Miklau, G. (2012). Pricing Aggregate Queries in a Data Marketplace. In WebDB (pp. 19-24).

Li, Y., Hou, M., Liu, H., & Liu, Y. (2012). Towards a theoretical framework of strategic decision, supporting capability and information sharing under the context of Internet of Things. Information Technology and Management, 13(4), 205-216.

Lin, I. C., & Liao, T. C. (2017). A Survey of Blockchain Security Issues and Challenges. IJ Network Security, 19(5), 653-659.

Lo, S. K., Xu, X., Chiam, Y. K., & Lu, Q. (2017, November). Evaluating Suitability of Applying Blockchain. In Engineering of Complex Computer Systems (ICECCS), 2017 22nd International Conference on (pp. 158-161). IEEE.

Matzutt, R., Hiller, J., Henze, M., Ziegeldorf, J. H., Müllmann, D., Hohlfeld, O., & Wehrle, K. (2018). A Quantitative Analysis of the Impact of Arbitrary Blockchain Content on Bitcoin. In Proceedings of the 22nd International Conference on Financial Cryptography and Data Security (FC). Springer.

Maymounkov, P., & Mazieres, D. (2002, March). Kademlia: A peer-to-peer information system based on the xor metric. In *International Workshop on Peer-to-Peer Systems* (pp. 53-65). Springer, Berlin, Heidelberg.

McKinsey report (2015), retrieved from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/creating-a-successful-internet-of-things-data-marketplace

Mengelkamp, E., Notheisen, B., Beer, C., Dauer, D., & Weinhardt, C. (2018). A blockchain-based smart grid: towards sustainable local energy markets. Computer Science-Research and Development, 33(1-2), 207-214.

Miller, K. (2012). "Big Data Analytics in Biomedical Research," Biomedical Computation Review

Mishra, A. K. (1996). Organizational responses to crisis. Trust in organizations: Frontiers of theory and research, 261.

Missier, P., Bajoudah, S., Capossele, A., Gaglione, A., & Nati, M. (2017, October). Mind My Value: a decentralized infrastructure for fair and trusted IoT data trading. In Proceedings of the Seventh International Conference on the Internet of Things (p. 15). ACM.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

Narayanan, A., Shmatikov, V. How to break anonymity of the netflix prize dataset. arXiv preprint cs/0610105, 2006.

Neuroni, A.C., Riedl, R., and Brugger, J., "Swiss Executive Authorities on Open Government Data - Policy Making beyond Transparency and Participation", IEEE, 46th Hawaii International Conference on Systems Sciences, 2013, pp. 1911-1920

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.

Ocean Protocol Foundation (2018), A decentralized Substrate for AI Data & Services. Technical Whitepaper
Parra-Arnau, J. (2018). Optimized, direct sale of privacy in personal data marketplaces. Information Sciences, 424, 354-384.

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data.

Perrin, B., Naftalski, F., and Houriez, R., "Cultural behaviour and personal data at the heart of the big data industry," E&Y and Forum D'Avignon 2013.

Peterson, J., & Krug, J. (2015). Augur: a decentralized, open-source platform for prediction markets. arXiv preprint arXiv:1501.01042.

Porter, M. E., & Millar, V. E. (1985). How information gives you competitive advantage.

Qian, H. (2018). Research on Data Security Storage Strategy in Cloud Environment.

Ramsundar, B., Chen, R., Vasudev, A., Robbins, R., & Gorokh, A. (2018). Tokenized Data Markets. arXiv preprint arXiv:1806.00139.

Rifi, N., Rachkidi, E., Agoulmine, N., & Taher, N. C. (2017). Towards using blockchain technology for eHealth data access management. In Advances in Biomedical Engineering (ICABME), 2017 Fourth International Conference on (pp. 1-4). IEEE.

Rifi, N., Rachkidi, E., Agoulmine, N., & Taher, N. C. (2017). Towards using blockchain technology for IoT data access protection. In Ubiquitous Wireless Broadband (ICUWB), 2017 IEEE 17th International Conference on (pp. 1-5). IEEE.

Roman, D., & Stefano, G. (2016). Towards a Reference Architecture for Trusted Data Marketplaces: The Credit Scoring Perspective. In Open and Big Data (OBD), International Conference on (pp. 95-101). IEEE.

Schilit B. N. and M. M. Theimer, "Disseminating Active MapInformation to Mobile Hosts," IEEE Netw., vol. 8, no. 5, pp. 22–32, 1994.

Schmid, B. F., & Lindemann, M. A. (1998). Elements of a reference model for electronic markets. Paper presented at the Proceedings of the Thirty-First Hawaii International Conference on System Sciences, Grand Wailea, Hawaii.

Schmidt, K. (2002) "The Problem with 'Awareness'." Computer Supported Cooperative Work: The Journal of Collaborative Computing, 11(3-4), pp. 285-298.

Schneier, B. (2015). Data and Goliath: The hidden battles to collect your data and control your world. WW Norton & Company.

Schwab, K., Marcus, A., Oyola, J. R., Hoffman, W. and Luzi, M., "Personal data: The emergence of a new asset class," 2011.

Sekaran, U. (2003). Research methods for business: A skill-building approach. New York: John Wiley & Sons.

Smith, G., Ofe, H. A., & Sandberg, J. (2016, January). Digital service innovation from open data: Exploring the value proposition of an open data marketplace. In System Sciences (HICSS), 2016 49th Hawaii International Conference on (pp. 1277-1286). IEEE.

Stahl, F., Schomm, F., & Vossen, G. (2014). The data marketplace survey revisited (No. 18). Working Papers, ERCIS-European Research Center for Information Systems.

Subramanian, H. (2017). Decentralized blockchain-based electronic marketplaces. Communications of the ACM, 61(1), 78-84.

Swan, M. (2015). Blockchain: Blueprint for a new economy. " O'Reilly Media, Inc.".

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Szabo, N. (1997). Formalizing and securing relationships on public networks. First Monday, 2(9).

Tapscott, D., & Tapscott, A. (2016). Blockchain revolution: how the technology behind bitcoin is changing money, business, and the world. Penguin.

Tene, O., & Polonetsky, J. (2012). Big data for all: Privacy and user control in the age of analytics. Nw. J. Tech. & Intell. Prop., 11, xxvii.

Thomas, L. D., & Leiponen, A. (2016). Big data commercialization. IEEE Engineering Management Review, 44(2), 74-90.

Tranfield, D., Denyer, D., and Smart, P., "Towards a methodology for developing evidence-informed management knowledge by means of systematic review," British Journal of Management, vol. 14, pp. 207–222, 2003.

Truong, H. L., Gangadharan, G. R., Comerio, M., Dustdar, S., & De Paoli, F. (2011). On analyzing and developing data contracts in cloud-based data marketplaces. In Services Computing Conference (APSCC), 2011 IEEE Asia-Pacific (pp. 174-181). IEEE.
US Government Accountability Office (2013) Consumer Privacy Framework Needs to Reflect Changes in Technology and the Marketplace; report to committee on commerce, science & transportation

Van Bommel, P., Van Gils, B., Proper, H. A, Van Vliet, M., and Van Der Weide, T. P., "The information market: Its basic concepts and its challenges," in 6th International Conference on Web Information Systems Engineering, WISE 2005, November 20, 2005–November 22, 2005, New York, NY, United states, 2005, pp. 577–583.

Van Engelenburg, S., Janssen M., Klievink, B. (2018) Designing context-aware systems: a structured method for understanding and analysing context

Van Engelenburg, S., Janssen, M., & Klievink, B. (2017). Design of a software architecture supporting business-to-government information sharing to improve public safety and security. Journal of Intelligent Information Systems, 1-24.

Van Niekerk, M., Van der Veer, R. (2017) Databroker DAO whitepaper: Global market for local data DAO.

Van Schalkwyk, F., Willmers, M., & McNaughton, M. (2016). Viscous open data: The roles of intermediaries in an open data ecosystem. Information Technology for Development, 22(sup1), 68-83.

Wong, J. I., & Kar, I. (2016). Everything you need to know about the Ethereum 'hard fork'. Qz. com, July 18.

Xia, F., Yang, L. T., Wang, L., & Vinel, A. (2012). Internet of things. International Journal of Communication Systems, 25(9), 1101-1102.

Xu, X., Weber, I., Staples, M., Zhu, L., Bosch, J., Bass, L., ... & Rimba, P. (2017, April). A taxonomy of blockchain-based systems for architecture design. In Software Architecture (ICSA), 2017 IEEE International Conference on (pp. 243-252). IEEE.

Yang, C., Chen, X., & Xiang, Y. (2018). Blockchain-based publicly verifiable data deletion scheme for cloud storage. Journal of Network and Computer Applications, 103, 185-193.

Zhu, H.W., and Madnick, S. E., "Legal challenges and strategies for comparison shopping and data reuse," Journal of Electronic Commerce Research, vol. 11, pp. 231–239, 2010.

Zikratov, I., Kuzmin, A., Akimenko, V., Niculichev, V., & Yalansky, L. (2017, April). Ensuring data integrity using blockchain technology. In Open Innovations Association (FRUCT), 2017 20th Conference of (pp. 534-539). IEEE.

Zuiderwijk, A., Loukis, E., Alexopoulos, C., Janssen, M., & Jeffery, K. (2014, May). Elements for the development of an open data marketplace. In Conference for E-Democracy and Open Governement (pp. 309-322).

Zyskind, G., & Nathan, O. (2015, May). Decentralizing privacy: Using blockchain to protect personal data. In Security and Privacy Workshops (SPW), 2015 IEEE (pp. 180-184). IEEE.

Appendix A: Interview protocols

Interviews
As is the case we are mainly involving businesses, and as business is largely a social phenomenon, it is important to conduct interviews as part of the insights that must come from practitioners.

The interviews conducted in this research are semi-structured, as there is a list of questions and areas that need to be discussed but the interviewer may ask follow-up questions or discuss another topic when he feels it is appropriate. Semi-structured interviews can provide reliable and comparable qualitative data.

Interview 1: Decentralized IoT sensor data marketplace provider

Settings
Date: The interview occurred on Wednesday May 16th 2018 in Leuven (Belgium), at Settlemint office.
Interviewer: Raphael Hannaert
Interviewee: Roderik van de Veer, CTO at Settlemint
Spectator: Cassandre Vandeputte, Business analyst at Settlemint
Interview recorded: Yes, after asking for permission
The interviewees were aware of the purpose of the interview but were not given the questions in advance.

Interviewee profile
Roderik van de Veer is CTO of Settlemint, a Belgian company which is working (among others) on the *DatabrokerDAO project*, a decentralized data marketplace for IoT sensor data using blockchain technology.

He is relevant for several parts of this research:

- He can provide his views on the flaws of current data exchange systems and how he is trying to solve these. For this first part, open questions were asked, such as about the problem statement, the type of data, the providers and users.

- He can give an overview of the technologies used for such platform, and especially justify the use of blockchain technology. The second part is therefore about the relevance of some technology decisions that the firm took, and some remaining technical challenges.

- Finally, he knows from his experience what are the adoption barriers for the technology, and we can research on how to overcome the barriers using the right design, when possible.

- He can also provide us with insights on what he believes is important to take into account when building and maintaining the marketplace.

Questions
The first part consists of open questions. For the second part, different values were proposed, with a short explanation of the value when the interviewee seemed not sure about the specific meaning of the value (as these can be quite broad). The interviewee had to judge each value on a numerical scale (see table below) and provide a justification.

Area of interest	Questions	Answer format
Current data exchanges	*-What is the background of DatabrokerDAO?* *-Which problem are you solving?*	Open explanation
Data and stakeholders	*-Is the data exchanged on the platform sensitive?*	Boolean + Open explanation
Technology, Blockchain	*-What technology choices have you made?* *-Why do you need blockchain technology?* *-How to make sure that data is not copied?*	Open explanation
Adoption and use	*-What are the remaining barriers for people to adopt the tech?*	Open explanation
Insights into the context	*-To what extent are the following values relevant for the design according to you:* *-privacy* *-efficiency* *-portability* *-interoperability* *-fairpricing*	Numbers on a numerical scale from 1 to 5, representing the five points in the following order: 1 = Not important at all 2 = Not very important 3 = I do not know 4 = Important

	-accessibility- *-censorship-resistance*	5 = Crucial The numbers are followed by open explanations

Interview 2: Data broker / decentralized data marketplace provider

Settings
Date: The interview occurred on Tuesday January 23rd 2018 in Singapore
Interviewer: Raphael Hannaert
Interviewee: Chirdeep Singh Chhabra, CEO of DEX, Co-founder of Ocean Protocol
Interview recorded: Yes, after asking for permission
The interviewees were aware of the purpose of the interview and was given the questions in advance.

Interviewee profile
Chirdeep Singh Chhabra is the CEO of DEX, an online marketplace platform for people (mainly organizations) to sell their data. After leading DEX for several years, he has also founded the subsidiary Ocean Protocol Foundation which aims at developing a decentralized data marketplace.

He is relevant mainly for the very first phase of the research i.e. understanding the background of decentralized data marketplaces such as the problem statement, the users of the product, the technologies used, the adoption, and applications.

The questions were mostly open and quite broad questions to get more information about the areas of interest illustrated in the table below.

Area of interest	Questions	Answer format
Problem statement	*Could you tell me more about your background and how you came up with the Ocean Protocol idea? What problem are you solving?*	Open explanation
Decentralization proposition and vision	*- How was the decentralization proposition accepted by your peers on the team? And by your clients and partners?*	Open explanation
Application, use-cases	*-What is for you an interesting use-*	Open

	case/industry for a decentralized data marketplace?	explanation
Technology	*You are working on a public blockchain. How do you see that in terms of scalability?*	

Questions
List of questions:
Could you tell me more about your background and how did you come up with the Ocean Protocol idea?
How was the decentralization proposition accepted by your peers in the team? And by your clients and partners?
You are working on a public blockchain. How do you see that in terms of scalability?
What is for you an interesting use-case/industry for a decentralized data marketplace?

The complete interview transcripts can be found at the end of this annex.

Interview 3 – Customer experience expert for businesses, working in a research-oriented technology consulting firm

Settings
Date: The interview occurred on Wednesday June 13th 2018 via Skype.
Interviewer: Raphael Hannaert
Interviewee: Brian Manusama
Interview recorded: No.
The interviewees were aware of the purpose of the interview but were not given the questions in advance.

Interviewee profile
Biran Manusama has more than twenty years of experience working closely with businesses and understanding their needs when using technologies. He is important to understand which elements businesses want when exchanging data, and what would restrict them from sharing.

Questions
The first part consists of open questions. For the second part, different values were proposed, with a short explanation of the value when the interviewee seemed not sure about the specific meaning of the value (as these can be quite broad). The interviewee had to judge each value on a numerical scale (see table below) and provide a justification.

Areas of interest	Questions	Answer format
Data sharing	To what extent businesses realize the potential of big data/data driven decision-making?	Open explanation
Data exchange mechanisms	How do they currently collect data? a. only internally, by for example doing a survey themselves b. via data brokers c. they may have a few contracts with other organizations d on open data marketplace e. on other data marketplaces	Multiple choice answer + Open explanation
Technology, requirements, context	What are the main points to consider in terms of customer experience when designing IT systems for businesses?	Open explanation
Context	What do you think would impact positively/negatively the willingness to participate to a data marketplace? (Privacy? User interface? Trust? Regulatory complexity?) As data providers (sellers in the marketplace) As data users (buyers in the marketplace)	Open explanation
Insights into the context	*-To what extent are the following values relevant for the design according to you:* *-privacy* *-efficiency* *-portability* *-interoperability* *-fairpricing* *-accessibility-* *-censorship-resistance*	Numbers on a numerical scale from 1 to 5, representing the five points in the following order: 1 = Not important at all 2 = Not very important 3 = I do not know

		4 = Important 5 = Crucial The numbers are followed by open explanations

Interview 4 – Data provider on a decentralized data marketplace

Settings
Date: The interview occurred on Wednesday June 19th 2018 via Skype.
Interviewer: Raphael Hannaert
Interviewee: Harm Van den Brink
Interview recorded: No.
The interviewee was aware of the purpose of the interview but were not given the questions in advance.

Interviewee profile
Harm van den Brink is a manager at ElaadNL, a dutch energy company. He is also part of the iota foundation, in charge of the governance and development of the directed acyclic graph based cryptocurrency iota. ElaadNL & IOTA have started the IOTA charging station and as part of this project cars and charging station exchange data and transactions autonomously on the iota decentralized data marketplace. Harm van den Brink is therefore a data supplier in our ecosystem.

Areas of interest	Questions	Answer format
Customers of sensor data marketplace	- What does your company do? -You have provided data on a decentralized data marketplace (i.e. iota data marketplace, charging station). Which data did you upload? - Who do you think could be interested by your data and for which purpose? - Are you planning to share more data?	Open explanation
Data exchange mechanisms	- Did you use to share data before such decentralized marketplaces exist? If yes, which mechanisms did you use?	Open explanation
Decentralized data marketplaces, problems of current exchange mechanisms	- Why did you choose a decentralized data marketplace, instead of current (non-decentralized) solutions such as Microsoft Azure Marketplace, Infochimps or data brokers?	Open explanation

Context	- What are for you some necessary conditions to find in a data exchange mechanism to convince you to share data? (e.g. monetary incentive, privacy)	Open explanation
Insights into the context	*-To what extent are the following values relevant for the design according to you:* *-privacy* *-efficiency* *-portability* *-interoperability* *-fairpricing* *-accessibility-* *-censorship-resistance*	Numbers on a numerical scale from 1 to 5, representing the five points in the following order: 1 = Not important at all 2 = Not very important 3 = I do not know 4 = Important 5 = Crucial The numbers are followed by open explanations

Interviews Transcripts

Interview 1

What is the background of databrokerDAO? Which problem are you solving?

What actually happened? We met with the innovation lab of Proximus and they said: well can't we use blockchain for something cool? This is what we have. We tried to see how to use blockchain to really make a difference. We figured out that there is a huge loss of potential in all the data streaming in their network. Why? Because they are doing some stuff with this data but could do more.

Apparently, it is allowed to use the cameras next to the road which scan the license plates. You can ask Proximus how many red Mercedes drive by on Monday mornings. They will not give the license plates, but they do some aggregate processing with the data that is flowing across the network, and we figured out that there is much more that you can do.
We researched deeper and then we came up with the content of the whitepaper. We worked on what could be the first offering. This is how it started.

Typically, the data that you deal with, in the sensors, are they sensitive?

No. I mean if you sell publicly it is a categorical no. For example, temperature [and] humidity; No personal data whatsoever. Now, we will not stop you from selling personal data. Therefore, as DatabrokerDAO we do not store nor see any data. It goes directly from you to the network, as a peer-to-peer transaction between you and the buyer.

There will also be a possibility to sell data to partners; to sell more sensitive data that would not be available publicly, more like a private sale of data.

We do not touch the stream between the sensors and the gateway operators (e.g. proximus). If you use a raspberry pi with a connection to proximus, the data will flow unencrypted. It is possible to encrypt the data as a sensor owner, but it then becomes much harder to sell. A coming feature that we will have is the proxy-encryption. Right now, we do not store but we will allow gateway operator to store historical data, because now when the sensor issues data once a day, and you buy it for 10 minutes you probably will not have anything. By having a historical data possibility such as last year, last month, we can have a better grasp of the data.

You are not dealing with personal data, so no GDPR compliance required?

Nothing at all. If the gateway operator would store information, or if data would be stored in the blockchain, then it becomes more difficult, but data is stored in a regular database.

Why do you need the blockchain tech?

What is the problem with selling data right now? If you look at the situation without a blockchain-type solution, the problem is that it's a lot of information, a lot of data that is existing and it is hard to contact each individual party directly and get contract with these parties, especially if we talk about gateway operators because they hate each other. They do not want to talk with each other and if you work with one probably you can't work with another one. The blockchain technology is used as an enabler to have the gateway operators to work together both globally and locally. Work together equals global markets for local data, by being fully decentralized we build a platform where no client of Proximus will ever see any other gateway operator name, but still have the advantage to buy data from providers for example in Dubai. Without the blockchain, which is independently possible to be read by everyone, you would have to become a big centralized party at the middle and build up trust which is much more difficult as many parties depend on you. By decentralizing and transmitting the data out of each gateway operator, data is flowing over a decentralized system. Even if you store data, it is the blockchain which provides the glue between the open-source API tool that we distribute as our product. The blockchain has the registration of the purchases, some kind of quality control, and token curation registry for the listing of sensors. It is interesting for the token economics part: to list a sensor, you have to stake coins, there is a minimum stake. As there is no intermediary to check the quality, we need a mechanism to convey confidence in data quality to convince data users to participate. The baseline of the story is put your DTX where your mouth is. If you want to list a sensor you have to lock up DTX. If you do not deliver what you promise, somebody can open a claim. To open a claim, you also have to stake, in order to prevent false claims. You can have conflict resolution via an arbiter or a voting system where everybody is an arbiter. If it is public, we have to allow everyone to see the data to vote, and thus the data loses value. The other solution is a third-party to resolve the conflict, that we will do in the first place. As soon as a claim is opened, it will appear in the interface, meaning that even if the claim is false you have a reputation

damage for a while. This is why it is important to stake when claiming also. In addition, the sensors are sorted based on the stake as well, on the website. Two sensors providing the same data will thus be sorted accordingly to the stake. In fact, more stake implies more certainty in the system. This is an economic system, not only from the goodwill of our heart.

Streaming sensor data, the data flows into the distributed API and sees who has bought it and who has been given access to the API, or datasets where we stored the encrypted information that can be downloaded. There, the staking is important as well cause if I say that have an excel with 10,000 people, I cannot deliver only 1,000 people. You need to be able to claim that you did not get what you are looking for.

We did not take the route of replacing people in the ecosystem. There will be 1.2 trillion sensors in 2019, which will not be replaced because you come with a nice blockchain technology. Some others try to replace gateway operators, but then it will mean convincing each sensor owner to not use these gateways. Instead, we just add an extra layer, a monetization layer, which benefits all parties in the system so nobody is against it, there is no conflict with any current stakeholders. We take a very targeted position, focusing on being an add-on. We are not replacing a whole ecosystem like other projects are doing.

We also aim at incentivizing people to participate in the system by processing the raw data and clean, aggregate, or put it through whatever big data or machine learning process, and then put it again in the data marketplace. The potential is then much bigger considering these leverages in making the data more valuable. In addition, it commoditizes the access to data. Data is already around, but if I want the temperature in Leuven I need contracts with all companies which have the sensors. With the marketplace, if any actor wants an aggregate of these data, it is possible as some data scientists in the network will provide these kinds of services, looking for data here and there and applying big data techniques in order to provide insights into these raw data. The output would thus be a dataset specifically tailored for an application. The current paradigm is more of a "push" type, with people uploading data on the portal and then the data requester search among these datasets. Ultimately, it could become more of a "pull" type, with the data requester broadcasting an information request, and data scientists buying and manipulating raw data to provide the information. It's like a network for big data and AI people.

How to make sure that data are not copied?

There is no way to avoid that. If you can see it, you have it. There is no way I can control what you do with. We believe that IoT data is time-sensitive. Value decreases over time. If somebody buys your data and sells it, you don't lose anything compared with now cause right now nobody is selling the data. So even if they share it, at least you have sold it once to them. If that would become a problem, we can always introduce terms & conditions with a contract. But the technology itself can not solve this.

What are the remaining barriers for people to adopt the tech?

The crypto parts, the tokens, where do you buy them, how does it work. I think that gateway operators will facilitate the process for users. They will do the staking, buy the tokens etc., but even the companies do not buy the tokens currently, we have to work with invoices.

Market share is important also for the success of the platform, you need sufficient sensors.

Either we run this on a private network with a bridge to use the tokens, which removes the ether from the equation part. Or we put it in the mainnet but then it makes it harder for enterprises. I think that we should put it in the mainnet and go from there.

Can you give me a level of relevance on a scale of 1 to 5 for the following values in your project? Where 1 = not relevant at all, 2 = not really relevant, 3 = I do not know/neutral, 4 = somewhat relevant, 5 = crucial

Privacy
Not very relevant as we are not selling personal data. However, there is some relevance for privacy when it comes to which business is buying which data, as it can give information about their strategy.

Efficiency (speed, throughput)
Somewhat relevant, but we are still far from reaching the maximum capabilities of the Ethereum network

Security (data not stolen)
Very relevant, but also very difficult. How do you know that somebody is not stealing your data flowing on the Proximus network?

Portability (e.g. mobile phones)
Less relevant, most interaction with the marketplace will be API bases. You want to buy million sensors; it's not something you do on the phone. More on the dataset level can be more interesting. But anyway, you do not buy cryptos on your phone, you need Metamask or so.

Interoperability (with other data marketplaces or blockchain technology)
Relevant. we try to not stop it. But not as relevant as something we need to do right now.

Fair pricing of the data
Not very relevant to implement rules ourselves. We consider that the market will figure it out.

Accessibility (everybody can access the product)
Very important

Censorship-resistance
Not relevant. We are totally not censorship-resistance. You can sell it, but all gateway operators are regulated entities, so if the government in Belgium says you cannot sell temperature data, then they cannot. Would that mean much more work telecommunications companies? Yes, to some extent.

End of transcript 1

Interview 2 - Transcript

Could you tell me more about your background and how did you come up with the Ocean Protocol idea?

Following my master in distributed systems at Ecole polytechnique fédérale de Lausanne (EPFL), I worked at IBM research labs and later at ETH, Zurich, in what people now call the Internet of Things. Later, I studied at the London Business School and worked in multiple ventures in London, most often in the data field. At this point, in my environment there was no doubt about the potential of data anymore, it had become commonly accepted that it was greatly valuable for businesses. The questions had moved to how to create value from the data, how do we unlock their potential. Finally, I joined DEX and moved to Singapore which has the ambition to become the first smart city. DEX had thus started working with the government and several enterprises here to build a centralized marketplace for 4 years now.

Why do we need a decentralized data marketplace?

One of the main problems in AI is the access to data: many AI companies came to me to connect them with people and organisations having these datasets. Actually, only a few companies have both datasets and machine learning algorithms (e.g. Facebook, Google). This is why we need some kinds of marketplaces to get access to data and enable transactions to happen. When I joined, I quickly realized that a centralized model was unable to scale. This is explained by the fact that entities would not give us their most valuable data for the simple reason that they cannot see what happens and then may feel that they lose control of their data. Transparency and ownership of data are important factors that complement the need for privacy and security of the data. Not having these characteristics fully operational was one of the biggest barriers before for our centralized exchange. For every single dataset, we should always know who the author is, almost like with citation in the academic field. Some are not satisfied with only a financial reward, which they may also lose if they are not officially owners anymore."

As a consequence, I started to look at alternatives and especially how blockchain technology and tokens can contribute. I have known Trent McConaghy (founder of BigchainDB, co-founder of Ocean Protocol foundation) for a while so I contacted him in Berlin. I told him about the idea of data being converted into assets that are traded within a tokenized ecosystem. Trent was writing articles about that and we shared the same view so we ended up creating Ocean Protocol, together with other members.

How was the decentralization proposition accepted by your peers in the team? And by your clients and partners?

Within the team we are all very optimistic about it and believe that this complete change in direction is necessary. This new philosophy ensures that Ocean Protocol is built in the right way, with a network of marketplaces upon it. This is a design that is important for the development of safe and sustainable AI.

Concerning the second part of the question, we have been discussing that with many of our clients. Actually, last year we had a large workshop with a number of C-level executives, and Data and Privacy Officers, about data management and sharing. They understand the value of data but problems appear when it comes to understanding the mechanisms of data access, regulations and compliance. They must be able to provide a list of who accesses the data upon request by regulators. Transparency and immutability are important factors that complement the need for privacy and security of the data. As I said, not having these characteristics fully operational was a big challenge before for DEX but there was much

enthusiasm when we elaborated on the decentralization, trust frameworks and the Ocean Protocol proposition. Convincing companies that are already working with data to join has logically been relatively straightforward.

We also have meetings with other corporates, not traditional data companies, those producing data on a daily-basis but not using them. We try to convince them of the need for allocating more resources in AI, data analysis/business intelligence. As an example, firms need to predict both the supply and demand (e.g. if some types of crops will grow in the coming years, or the consumption of end-products). In addition, even if they produce more and more data, this is not enough to have accurate forecasts and stay competitive. They need external data for a rich insight and forecast. That's why we need the marketplace where they can buy and sell data (it can also create new revenue streams) in order to complement the data they are producing. My conclusion is that if companies do not participate in the data markets they will be excluded from the future data economy and may be at risk of shutting down.

You are working on a public blockchain. How do you see that in terms of scalability?

BigchainDB has built a scalable blockchain database. We have a history around that. Nevertheless, we understand that there are technical challenges and therefore we need to partner with other projects and scientists but as soon as possible also with the community using the open-source protocol.

In terms of product development, we aim at coming with a first Minimum Viable Product by Q3 2018 and network launch by Q1 2019.

"The General Data Protection Regulation was designed to harmonize data privacy laws across Europe, to protect and empower all EU citizens data privacy and to reshape the way organizations across the region approach data privacy. Approved on 14th April 2016, it will be enforced on 25th May 2018 at which time organizations in non-compliance will face heavy fines". Source: https://www.eugdpr.org

Does that mean that I could also sell my data?

Nothing would prevent you of doing that. However, at the beginning you will have no credibility on the network, so you would need to be referred or put stake (Note: put money at stake means buying and betting tokens such that if one's data appeared to be false or not actually her, that person will lose her stake (and could even be blacklisted), quite similar to how proof-of-stake achieves consensus in some public blockchains). This is why at the beginning we are starting with those that have larger and valuable datasets. Nevertheless, we are building the token economy with in mind the purpose of not allowing any kind of centralization so of course it will be possible.

How do we ensure quality?

The rewards that one gets as a result of his data being very popular is logarithmic. Therefore, you cannot take over control as there are incentives for people to work with new

data (because of the logarithmic curve). This mechanism ensures that people work, curate and bring new data. Price will have probably little to do with the popularity of the data. In any case, it is not our job to attach that. Data providers have the right to judge which price to set and rules are defined for the data marketplaces by keepers.

It is also important to understand that policies can change depending on the marketplace as they can be subject to different regulations and purposes. There may be some marketplaces specific to some fields like healthcare and energy. As stated previously, we do not think that there will be only one global marketplace.

What is for you an interesting use-case/industry for a decentralized data marketplace?

In my opinion, the most impactful one is healthcare. As an example, in the context of the Parkinson disease some companies are working on AI application to define the right scale of accuracy for tremors measurements. This input is then used to estimate the right dosage, duration and how often patients need to take the medicine. If the condition is not managed properly, they may need to have an implant in their brain which costs about 50.000€. This is a very expensive operation that more accurate machine learning predictions could replace. However, to get a low error rate, we would need 10.000 patient's data. It is clear that no hospital can provide such amount of data, but a decentralized data marketplace can. Thanks to distributed ledger technologies, the sharing of patient data will be enabled but data will still remain with the patient or within the hospital. An algorithm that has been developed in Singapore could be sent to a hospital in Munich (after making sure that the data are formatted accordingly) for training and returns to Singapore without bringing back data. Moving algorithms is cheaper than moving data. We just need to prove that no data is pulled, which we believe is not difficult to achieve. In this case compliance and regulation are satisfied, the AI is trained, and the impact is happening.

End of transcript 2

Interview 3 – notes

I believe businesses no business are not seeing the impact of data driven decision, no one does not see the potential.
sales, customer up sell cross sell, what do i need to cross sell.

currently collecting?
huge difference between business does and what IT does, business side know and say we need "everything" the result of that is that the IT department has huge data warehouses with all sources of data but not targeted at all to business purpose to take better decision. If you have 100% of business that understand but only small % which is actually doing this.
Holy grail: acquire all data, normalize all data, extract customer behavior from it, understand the journey which is associated and how customer is working with us, predictive model to see what would be the next action. Potential they are turning, how do i need to respond, which engagement channel. A lot of data points which I am filling my data lakes with.

How many companies do actually do this? 1%

collecting a lot of data: 70%
hard to understand,
Customer experience = competitive advantage in marketplace, not product anymore therefore need to understand behavior, so need to understand the current experience, and influence the behavior. We are just trying to understand who the customer is, not yet engaging into proactive strategy.

Not a lot engaging with data from other companies, dealing with their own mess first. Artificial Intelligence: machine learning need to have large amounts of data to ingest into the machine learning engine, thats where a lot of organizations are lagging data at the moment, interested in marketplaces for this purpose. Data marketplace is the catalyst for AI.

IBM watson problem: not a lot of customers, not a lot of data, empty conversational platform capability. The dont have data from the customers.

More contracts where company asks for using data in an anonymized way in exchange for a discount, so that they can use the anonymized data. Increase their understanding and training with the data.

Share data with me:

financial compensation
benefit from data from competitors, well trained machine learning. Does not explain how they are executing. Competition will be on execution, not on understanding. Sales power, structure of the organization to take advantage of the insights. training set to have a machine learning to reach a higher accuracy.

Huge confusion going now on the market because of the GDPR, people are very reluctant to use, sell, given consent for using the data because of fear. GDPR is hurting innovation, shock reaction. Companies very rigid on what to do. Not anymore going to the cloud to buy and sell. Companies are very afraid.

companies had mentality to sell data without knowing where it would end up. Thats why we have GDPR. GDPR is giving us the option to select upon who we are selling the data to because of responsibilities. Would be liable.

interesting marketplace: not only about customer experience but things like cancer treatment, find the best cure.

End of notes interview 3