IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 15, 2022, accepted March 12, 2022, date of publication March 22, 2022, date of current version March 30, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3161478

Data Market Design: A Systematic Literature Review

STEFAN W. DRIESSEN, (Graduate Student Member, IEEE), GEERT MONSIEUR, AND WILLEM-JAN VAN DEN HEUVEL

Jheronimus Academy of Data Science, Tilburg University, 5211 DA 's-Hertogenbosch, The Netherlands

Corresponding author: Stefan W. Driessen (s.w.driessen@jads.nl)

This work was supported in part by the European Union Horizon 2020 Project under Grant 825480.

ABSTRACT Data markets are platforms that provide the necessary infrastructure and services to facilitate the exchange of data products between data providers and data consumers from different environments. Over the last decade, many data markets have sprung up, capitalising on the increased appreciation of the value of data and catering to different domains. In this work, we analyse the existing body of scientific literature on data markets to provide the first comprehensive overview of research into the design of data markets, regardless of scientific background or application domain. In doing so, we contribute to the field in several ways: 1) We present an overview of the state of the art in academic research on data markets and compare this with existing market trends to identify potential gaps. 2) We identify important application domains and contexts where data markets are being put into practice. 3) Finally, we provide taxonomies of both design problems for data markets and the solutions that are being investigated to address them. We conclude our work by identifying common types of data markets and corresponding best practices for designing them. The outcome of this work is intended to serve as a starting point for software architects and engineers looking to design data markets.

INDEX TERMS Data market, data marketplace, data product, literature review.

I. INTRODUCTION

Nowadays, data is no longer viewed as an inept byproduct of (business) processes, but rather a valuable resource [1], [2]. A famous analogy proclaims data as the new oil,[1] and, like oil, it can be traded, processed and used in different contexts and applications. Indeed, the last decade has seen an incredible increase in both the amount of data being collected [3], [4], as well as the development of infrastructure necessary to process and share the vast amounts of collected data in new contexts [5], [6].

In the wake of these trends, many data markets have sprung up, facilitating data exchange between data providers and data consumers. These data markets capitalise on the increased appreciation of the value of data, catering to different domains (e.g., IoT [7], medical data [8] manufacturing data [9]) and contexts (e.g., national data [10], [11]). Therefore, it is not surprising that the scientific community has taken an interest in the phenomenon of data markets as well: as fig. 1 shows, there is a definite trend in scientific articles being published that have a term related to data market(places) in their title or keywords. In this work, we analyse the existing body of scientific literature on data markets to provide the first

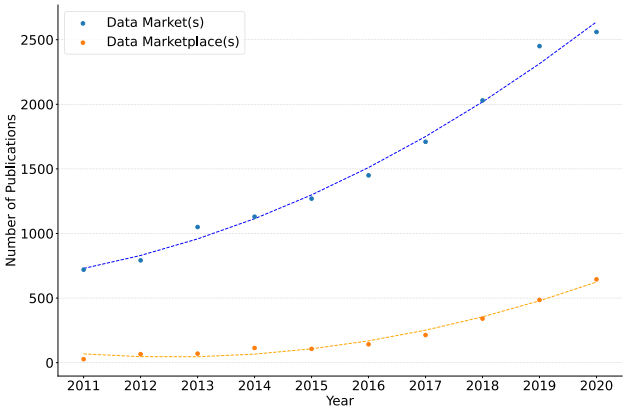


FIGURE 1. Research Trends for Data Markets, an exponential growth is observed. Source: Number of results for each query in google scholar.

The associate editor coordinating the review of this manuscript and approving it for publication was Kostas Kolomvatsos.

[1]The Economist, "The world's most valuable resource is no longer oil, but data," may 2017

comprehensive overview of research into the design of data markets, regardless of scientific background or application domain.

A. PREVIOUS SURVEYS ON DATA MARKETS

Several literature reviews and surveys have been conducted to review and discuss both academic literature and industry developments on data markets. Particular interest has come from the field of business management, where reviews of real-world data markets (and sometimes academic work) aim to provide organisational and business insights. In this category, Thomas and Leiponen provide one of the first type of reviews; they distinguish six basic business models that emerged in commercial data markets up until 2013. Furthermore, they observe several key characteristics of the ecosystems in which these business models are applied and conclude with the main challenges for commercialising data [12]. Stahl *et al.* also take a business management view but focus more on the data market itself rather than the business models associated with it. They use neoclassical economics to arrive at a formal definition of a data marketplace and present a classification framework to distinguish between six different types of data markets across three dimensions: business model, ownership, and hierarchy [1]. Similarly, Koutroumpis *et al.* focus on data markets and consider how markets for data differ from traditional digital markets by considering the economic and technical properties of data products. They conclude that proving data provenance is essential for the success of data markets and finally propose a classification of data markets that is very similar to that of Stahl *et al.* [13]. Finally, there exist some surveys such as the series of reports by Stahl *et al.* [14], and the work of Spiekermann [15]. These surveys look at real-world data markets in Europe, the U.S., and Israel to identify industry trends and reinforce the classification frameworks from the works above.

Several other literature reviews have come from the domains where data markets are being developed and implemented, particularly from Internet-of-Things (IoT). For example, Ishmaev provides an extensive overview of the limitations of blockchain technology for trading personal IoT data, such as data resulting from smart homes or wearable sensors [16]. Further insight into the societal application of data markets for IoT is provided by Barns, who classifies and evaluates platforms that different governments have developed to extract value from their data sharing efforts [17]. Finally, Brandão *et al.* provide a general (unstructured) literature review of IoT data markets, identifying lack of trust and reliability of data providers as the main problem in this context [18]. In addition to IoT data markets, particular attention has also been given to different methods for protecting the privacy of individuals whose data is shared. In this context, Perera, Chang *et al.* give an overview of thirteen different privacy modelling languages that aim to express information about privacy and its management [19]. They conclude that data markets come with additional requirements that existing privacy modelling approaches cannot yet meet. Meanwhile, Perera, Wakenshaw *et al.* discuss the benefits and challenges of adopting "data boxes" [20]. Data boxes are personal devices for storing personal data that allow the owner to control who gets access to which data. Finally, ethical challenges of trading personal medical data are discussed by Ahmed & Shabani, who, in the context of DNA data, note that it can be especially challenging to communicate to data providers the limitations and possibilities of sharing their sensitive data [8].

Finally, out of all the different challenges for data markets, *pricing mechanisms* attract perhaps the most attention from the academic community. Several literature reviews exist that compare different ways of pricing data. For instance, Muschalle *et al.* interviewed seven established data vendors back in 2013 and present insights regarding their pricing strategies, noting along the way that these strategies are changing rapidly. Liang *et al.* define a life cycle of data in the context of data trading, categorise various pricing mechanisms and discuss how these affect challenges such as security and maintaining data sovereignty [21]. Golrezaei *et al.* discuss different pricing schemes applicable to data markets for traffic data, paying particular attention to the differences between 'raw' sensor data, which they show to be less valuable, and 'processed' data, which is perceived as more valuable [22]. Finally, Fricker *et al.* performed a structured literature review with snowballing and provide the most extensive overview of pricing data products in the academic literature on data markets [23]. They identify fourteen different approaches and evaluate each of these on functionality, context and maturity.

B. MOTIVATION AND CONTRIBUTIONS OF THIS SURVEY

Despite the works discussed above, it has been made clear that data market development suffers from a lack of common definitions, design standards, and terminologies [1], [24]–[26]. We speculate that this is partly due to the fact that all meta-research mentioned above has focused either on organisational aspects of running data markets, business aspects, such as how to price data products, or focused exclusively on one domain or feature of data markets (e.g., IoT or pricing strategies). What is still lacking, however, is a perspective that considers data markets as IT artefacts with *technical implications* with manifestations *across different domains and contexts*. We believe such a perspective could prove valuable to engineers, managers and entrepreneurs alike, who are setting out to *design* and *build* a data market by providing insights on how to go about doing that. Therefore, this paper explores the challenges and solutions for designing data markets by contributing a structural literature review of data markets in a holistic, cross-domain perspective. In particular, we attempt to go beyond existing perspectives and also consider the technical challenges for designing a data market. In doing so, we contribute to the field in three distinct ways:

1) We present an overview of the state of the art in academic research on data markets and compare this with existing market trends to identify potential gaps.

2) We identify important application domains and contexts where data markets are being put into practice. Moreover, we consider how these domains and contexts affect the possible architectural designs and manifestations of data markets. Additionally, we identify the concepts that persist across different domains and arrive at a minimal definition of a data market.
3) Finally, we provide taxonomies of both design problems for data markets and the solutions that are being investigated to address them.

The rest of this article is organised as follows. Section II discusses the typical concepts considered in all formal and informal definitions of data markets before coming to a minimal definition of a data market that can be applied to all the works investigated for this literature review. Next, section III introduces the methodology that was applied in our review. Section IV presents the main results of this study. Based on the results Section V describes a set of types of frequently occurring data markets and their best practices applied in the design of these markets. Finally, Section VI summarises and discusses the main contributions of the present study.

II. BACKGROUND AND BASIC CONCEPTS

Data markets come in different shapes and sizes, and different authors have provided different definitions of the term data market or data marketplace. For example, some authors such as Sharma *et al.*, Stahl *et al.* & Spiekermann consider a data market to be a version of a digital market such as ebay[2] that specialises in monetising data [1], [15], [27]. Other definitions disregard the need for payment and focus simply on the act of exchanging data products in a way that is convenient for the actors involved and respects their needs [28], [29]. Finally, some definitions consider data markets to be a platform where new data utilisation- and value creation methods and are *"created through the process of interaction between"* different actors [30], [31]. Despite these apparent differences, four concepts can be considered part of all formal and informal definitions of data markets: 1) data products, 2) data providers, 3) data consumers and, 4) the action of exchange. We briefly introduce each of these concepts below before arriving at a minimal definition of a data market that can be applied to all the works investigated for this literature review.

A. DATA PRODUCT

In order to understand data products, it is necessary to introduce the concept of *data assets*, which is any digitally stored information that has potential value. Data assets are owned by a *data owner* who wants to make them available on the data market. When data assets are optimised for consumption by external data consumers on a data market, they become *data products*. Examples of efforts for this optimisation include standardisation of the data (e.g., with the help of data modelling), the inclusion of (standardised) metadata to describe the data asset, the creation of access & usage policies, creation of access points (e.g., through the use of API's), and the registration of the data product in some centralised registry to make it easier to discover the data product.

[2]https://www.ebay.com/

B. DATA PROVIDER

Data providers are those actors responsible for the creation, maintenance and general operation of data products. A data provider will often also be the owner of the data and, in some cases, the creator of the data, but this need not be the case. It is possible and, in some cases, practical that data products are created on behalf of the data owner [32]. Nevertheless, every data market presumes the existence of data providers as key actors.

C. DATA CONSUMER

Data consumers are the actors who consume data products that have been made available on a data market by data providers. An important thing to note is that the data consumers come from different environments compared to the data consumers, which is the main reason for the existence of data markets in the first place. These different environments can arise from different geographical locations or organisations, but they can also be different teams inside a large company that have little interaction with each other and require a data market platform to use each other's data assets.

D. EXCHANGE

The act of exchanging on a data market implies that both parties, the data provider and the data consumer, will gain and give something on the data market. On the one hand, data consumers gain value from the data product, which the data provider lets them consume somehow. On the other hand, the data provider might gain anything from money to new insights or improved data-driven processes or services from the data consumer.

Based on these concepts, we now provide a minimal definition of a data market, which fits all data markets investigated in this work:

> **A data market is a platform that provides the necessary infrastructure and services to facilitate the exchange of data products between data providers and data consumers from different environments.**

This definition is broader than the existing definitions discussed above and allows for manifestations that have been hitherto excluded by one or more of those definitions such as internal data markets [33], [34], data markets for public well-being [17], [35] and data markets where the data provider receives services and insights in exchange for their data [36]. For example, this definition of a data market allows us to consider some social media platforms (e.g., Facebook,[3] Linkedin,[4] Reddit[5]) as data markets. Their users

[3]https://www.facebook.com
[4]https://www.linkedin.com
[5]https://www.reddit.com

(data providers) provide data in exchange for the services offered by the platform. The data is valuable, and the platform provider uses it to offer enhanced services (advertisements) to their customers. The benefit of using such a broad definition is that, even though the manifestations of data markets are quite different, many elements are still the same (or similar) and the insights in this literature review carry over from to other manifestations.

III. RESEARCH METHODOLOGY

Systematic literature reviews aim to synthesise the results of many different contributions on a particular topic. This is valuable to practitioners, who often do not have time to keep up with all available research [37], as well as scientists, who aim to position their work in the context of existing literature and the state of the art [38], [39]. In this work, we follow a set of guidelines set out expressly for information systems research by Tranfield *et al.* [40] as well as Kitchenham and Charters [39]. Based on these guidelines, our systematic literature review is organised in three phases: the planning phase, the execution phase and the reporting phase. A high-level overview of our methodology and the three phases is shown in fig. 2. The rest of this section describes each of the phases and the steps that were taken in them.

A. PLANNING PHASE

Before conducting a systematic literature review, the first step is to identify its need. As was already argued in section I, there is a need for a systematic literature review that supplements the business management and organisational perspectives with a design-oriented one, and that draws from all work on data markets, regardless of their domain, to arrive at common definitions and terminologies for those design concepts that are not restricted to *one* specific domain. However, this does not mean that the difference in domains should be considered irrelevant. On the contrary, we hypothesise that different domains offer valuable perspectives on data markets through the roles, problems, and solutions they consider relevant for designing data markets.

In order to guide our literature review towards the desired insights and attain the three contributions mentioned in section I, we formulate four research questions that guide our investigation of the selected papers. These questions were inspired by the guidelines for developing research questions for systematic literature reviews presented by Kitchenham [42]. The first research question can be traced back to contributions one and two, the second research question corresponds to contribution two, and the last two research questions correspond to contribution three. We introduce these questions and discuss them briefly below:

RQ 1: Which types of data markets are being investigated in literature?

RQ 2: What are the application domains of data markets described in literature, and which main roles can be identified in these domains?

RQ 3: Which problems have been identified that hinder the successful implementation of data markets?

RQ 4: Which solutions are being proposed to solve these problems?

The goal of RQ 1 is twofold: on the one hand, it allows us to build on and connect our work to existing literature. As discussed below, we build on the existing classification frameworks proposed by Stahl *et al.* [1], Koutroumpis *et al.* [13], and Spiekermann [15] to identify how the data markets investigated from organisational or business management perspectives manifest themselves when viewed from a technical perspective, as well as identify research trends in academia. To this end, a natural sub-question is asked:

RQ 1.1: Which are the business management and organisational properties of data markets in academic literature and how do they compare to real-world data markets?

On the other hand, RQ 1 allows for an evaluation of *what* is being proposed, produced and evaluated by academic literature in the context of data markets. We thus consider which artefacts (i.e., the "things" under investigation) are being produced by academic literature and to what extent these artefacts are generalisable to practice by asking a second sub-question:

RQ 1.2: Which artefacts (e.g. a data market architecture) are being produced by academic literature and how often are these artefacts evaluated in a real-world scenario?

Since this review aims for a design-oriented approach to data markets, it takes inspiration from design science to guide its research questions. In the design science research approach, the context is considered systematically by looking at different roles and properties from the domain in which the solution is to be implemented [43], [44]. The second research question extends the first by providing an overview of the domains and contexts from which the academic works stem. Additionally, it is valuable for scientists, engineers, and managers looking to design a data market because it provides a starting point to existing knowledge. Therefore, two sub-questions are asked for RQ 2:

RQ 2.1: In which application domains are data markets being investigated academically?

RQ 2.2: Which roles are directly involved in the design, operation and maintenance of data markets?

In this context, we define the role of an actor by the responsibilities that they take on. That is, actors are entities who fulfil one or more roles by taking on the responsibilities associated with this role. This also means that the data provider and data consumer discussed above can be considered roles for actors responsible for providing and consuming data products, respectively.

Finally, by looking at implementations of data markets across all possible contexts, it becomes possible to arrive at key concepts of data markets and create a minimal definition of the term data market itself. Therefore, the final sub-question is:

RQ 2.3: Which are the concepts that are consistent across different definitions of data markets?

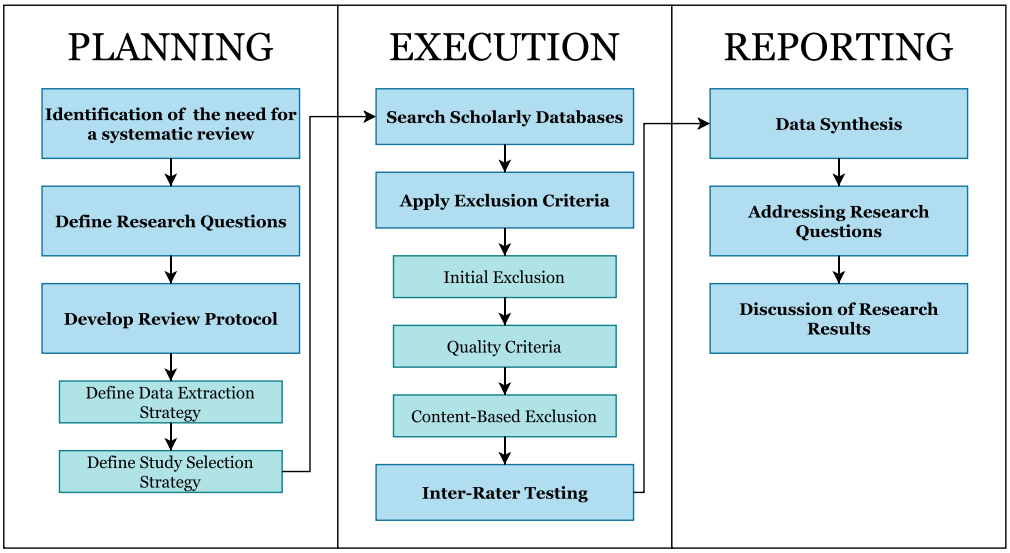


FIGURE 2. Overview of the systematic review methodology, adapted from [39]–[41]. The three phases of executing a systematic literature review are shown. Each step in the phase is shown in a box, with sub-steps shown in smaller boxes and arrows indicating the sequence in which the steps were taken.

Finally, RQ 3 and RQ 4 allow us to evaluate the direction of current research into data markets, make predictions about what type of research will be most useful in the coming years and how data markets could evolve in the future. These questions can be combined with RQ 2 to find contextual differences in the focus on problems or solutions. Moreover, the resulting taxonomies will help guide future design efforts for data markets by providing critical insights into the relevant problems and solutions.

1) DATA EXTRACTION STRATEGY

Following the methodology of Kitchenham and Charters [39], we define an extraction form based on a set of dimensions and corresponding values. The dimensions are chosen in order to address the research questions, and the extraction form provides the basis for the data synthesis step [40]. Table 1 shows an overview of the dimensions used for data extraction, the set of values that these dimensions can take, and the research question each dimension contributes to answering. Most of the possible values and dimensions were taken from existing literature (reviews) on data markets; however, for the 'Application Domain', 'Roles', 'Problems Addressed', and 'Solutions Proposed' dimension, the potential values result from applying grounded theory. Grounded theory is a systematic methodology for qualitative research with the ultimate goal of deriving theory from a large number of textual sources [45], [46]. This is achieved by successively merging and refining similar ideas captured in documents into categories that comprise theories. The process of applying grounded theory is further explained below in section III-C. First, the motivations behind each dimension and the possible values are explained below.

2) AUTHOR

The author dimension is tracked simply for administrative purposes and is used primarily to keep track of different papers.

3) YEAR

The year in which each article is published is beneficial not only for administrative purposes but also for identifying trends in research interest over the period investigated. As possible values, we only consider works published after 2015 to limit our selection, focusing our attention on state of the art and extending the existing literature reviews.

4) DATA MARKET DEFINITION

The data market definition tracks whether the work under investigation provides or cites a formal or informal definition. If it does, we note the full definition as it is written. Keeping track of this dimension allows us to discover whether different domains have different ideas on what a data market is and establish a minimal definition of a data market. The analysis of the different definitions has led to the background and basic concepts discussed in section II.

5) MATCHING MODEL

The matching model dimension comes from the business management perspective on trading and describes how matching between "buyers" (data consumers) and "sellers" (data providers) is achieved in a data market [47], [48] and has been used for previous classifications of data markets. Stahl *et al.* [1] consider six types of "business models," which are analogous to the four types of matching models discussed by Koutroumpis *et al.* [13]. We use the matching models rather than the business models because they are

TABLE 1. The dimensions considered for data extraction, their possible values and the (research) question they are intended to address.

Dimension	Possible Values	Question addressed
Authors Year	Names 2016-2021	Who published the article and when?
Data Market Definition	Full definition as written	What is the minimal definition of a data market?
Matching Model Platform Owner Platform Architecture Time Frame Data Access Method Artefact	One-to-one, One-to-many, Many-to-one, Many-to-many Private Seller, Private Buyer, Consortium, Independent Centralised, Decentralised Static, Dynamic Direct Download, Pull-Based, Push-based, Compute-to-Data, Specialised Software Data Market Architecture, Literature Review, Evaluation	Which types of data markets are investigated in literature? (**RQ 1**)
Application Domain Roles	AI & Machine Learning, Automotive & Mobility, Domain Agnostic, Industrial, Governmental Data, Healthcare, IoT, Personal Data, Smart City, Smart Home, Data Provider, Data Consumer, Platform Owner, Platform Software Provider, Data Broker, Clearing House, Data Transformer, Quality Assessor, Identity Provider, Certification Provider, Infrastructure Provider	What are the application domains and main roles identified in the work? (**RQ 2**)
Problems addressed	Data Brokering, Data Governance, Data Quality, Data Sovereignty, Data Transformation, Efficiency, Ethical Concerns, Maintainability, Performance, Price Determination, Reliability, Resource Minimisation, Scalability, Security, Strategic Behaviour, Transaction Enforcement, User Friendliness	Which problems are being addressed? (**RQ 3**)
Solutions proposed	Access Control Mechanism, Automated Data Transformation, Autonomous Actors, Certification Framework, Compute to Data, Crowd-Sourcing, Data Transformation Environment, Dispute Management System, Data Description Standard, Encryption, Legal Enforcement, Logging, Identity Management Mechanism, Manual Brokering, Manual Data Transformation, Origin Tracing Mechanism, Participant Management, Quality Metrics, Quality Assessor, Querying Mechanism, Recommender System, Reputation, Usage Policy Management	Which solutions are being proposed? (**RQ 4**)

more closely aligned with our goal of providing a technical perspective.

One-to-one matching occurs when a single seller sells to a single buyer based on a bilateral agreement. In the context of data markets, this can occur when data providers create data products that are tailored to the specific needs of the data consumers (e.g., the service offered by Data and Sons[6]). Another matching model is one-to-many, which occurs when a data provider sells the same data product to multiple data consumers; an excellent example of such a matching model is Twitter, which provides standard APIs for data consumers to query.[7] Many-to-one matching is akin to data harvesting, with one data consumer and many data providers: each data product goes to exactly one data consumer. An example of a data marketplace that is based on many-to-one matching is Facebook,[8] which collects data of its users in exchange for access to its social media platform. Finally, many-to-many matching corresponds to the scenario where multiple data providers exchange their data with multiple data consumers. Such a platform is especially interesting because it generally only *facilitates* data trading but does not take ownership of the data that is traded [1].

[6]https://www.dataandsons.com/request-datasets
[7]https://developer.twitter.com/en/docs/twitter-api
[8]https://www.facebook.com/

6) PLATFORM OWNER

The platform owner dimension comes from the work of Stahl *et al.* [1] and describes the relation of the actor who provides the data market to that of the data provider and data consumer. As will be discussed in section IV the question of who operates the data market has far-reaching consequences for the way a data market operates; for example: if the data provider is also the platform owner, it is easier for them to make sure all data is provided in a standard format and to set a price for the data product(s) provided [21], [23]. First, the quality of a data product has a service-level facet, which consists of extra-functional properties usually described in a service-level agreement (SLA), such as the availability of the data product and the responsibilities of the data provider and the data consumer [49]. Additionally, data quality has a content-based facet driven by the information contained in the data product.

Generally speaking, data markets that operate under a one-to-many or many-to-one matching model are owned and operated by the 'one' in the relation. In that case, the data owner can be considered a private seller or private buyer, respectively. Alternatively, a platform can be operated by a consortium, such as the BONSEYES initiative [50]. Finally, the platform owner may also be an independent party, which facilitates one-to-many (sell-side), many-to-one (buy-side) or

many-to-many (two-sided) data trading for data providers and data consumers other than themselves. The main difference between consortium-owned and independently-owned data markets lies in the level of control that the platform owner is able to exert over the participants in the network, as will be discussed in subsequent sections.

7) PLATFORM ARCHITECTURE

The platform architecture dimension was taken from Spiekermann [15] and describes two fundamentally different approaches to designing a data market platform: in a centralised approach, all data that is traded goes through a centralised (often cloud-based) point. Because of this, a centralised architecture enables easy standardisation of the data exchange, processing and access control management. Alternatively, a data market platform can be designed decentrally, putting the burden of organisation on many different roles, most often the data providers and data consumers themselves. The advantage of this approach is the control it gives to data consumers and data providers to protect their interests, such as data sovereignty and data governance.

8) TIME FRAME

The time frame dimension comes from the earlier classification efforts of Schomm *et al.*, and Stahl *et al.* [14], [24], [25]. This dimension concerns the *currentness* of the data, which can be either static, meaning it is factual data that is valuable for an extended period (e.g., an image set that can be used for training a recurrent neural network) or dynamic, meaning it will quickly lose value over time (e.g., real-time traffic data).

9) DATA ACCESS METHOD

The data access method determines the way the data consumer gains access to the data product. Like the time frame dimension, the data access method dimension has been adopted from the work of Schomm *et al.* and Stahl *et al.* [14], [24], [25]. They identified four potential values for this dimension: 1) API (application programming interface) provides programmatic access to a data product via the internet, 2) download, meaning a data set is transferred from the data provider to the data consumer, 3) specialised software means that the data market uses a custom software solution to facilitate data transfer and 4) web interface means the data is presented to the data consumer on a website. Based on our findings, some slight changes were made to these possible values. Firstly, for APIs, a distinguishment is made between *pull-based* and *push-based* access methods such as streaming. In a pull-based approach, the data consumer has to take the initiative by making requests to pull the data towards them, whereas in a push-based approach, the data consumer subscribes to a topic or a channel is created for streaming and has the data pushed to them by the data provider [51]. Moreover, the web interfaces we found were all significantly different from each other and often had significant overlap with specialised software, so these access methods were classified as using specialised software. Finally, a novel access method emerged from the literature review, namely, compute-to-data, whereby the data consumer is granted the ability to run code on the data and receive aggregate results without actually being able to access the data product directly [52].

10) ARTEFACT

Artefacts are the main objects under investigation in design science approaches [43]. Since our goal with this paper is to provide a design perspective, we keep track of whether the work proposes a literature review or overview on data markets, an entire data market architecture or a methodology for designing or implementing a particular aspect of a data market. Additionally, since we are interested in the generalisability of the results to industry, we keep track of whether the researchers worked together with a partner from industry to evaluate the artefact in a real-world scenario.

11) PROBLEMS ADRESSED

The problems addressed dimension keeps track of the different problems that are identified in the work considered and allows us to answer RQ 3. Following the approach of grounded theory, the values are initially codes but are later abstracted towards the concepts shown in table 1

12) SOLUTIONS PROPOSED

The solutions proposed dimension keeps track of the different types of solutions that are proposed in the work and allows us to answer RQ 4. Following the approach of grounded theory, the values are initially codes but are later abstracted towards the concepts shown in table 1

13) APPLICATION DOMAIN

The application domain is the domain in which the data market(s) discussed in each work operate and from which the researchers that propose the data markets themselves hail. That is to say; it is the domain from which the data providers, the data itself and often also the data consumers stem. Together with the identified roles, the application domain is the most important aspect of the context of the artefact when considered from a design science perspective [43].

14) ROLES

Stakeholders can be defined as a person, group of persons or institution affected by treating the problem that the artefact addresses [43] and, as such, have to be considered when designing a data market. In an attempt to capture the most important stakeholders, the roles of the actors who are are considered to participate in the data market actively are abstracted as concepts through the process of grounded theory. The roles, including a view on the actors who fulfil them, are crucial to understanding the problems and solutions as they are the ones that experience the problems and benefit from the provided solutions.

15) STUDY SELECTION STRATEGY

In order to arrive at a suitable selection for answering our research questions, we apply a study selection strategy which, following the protocols laid out by Kitchenham *et al.* [39], consists of three steps: (1) define a search string, (2) apply the search string on selected search engines and (3) extract relevant papers by applying pre-established exclusion criteria. Figure 3 outlines the study selection process and shows how many papers were included or excluded during each step.

The process of selecting a good search string is often based on trial searches, which are indicative of the usefulness of the applied search string [53]. For this systematic review, we started by manually selecting a small set of relevant papers on data markets and trying different search string combinations to see (1) whether they yielded a significant number of resulting works and (2) whether the results included our manually collected set of relevant papers. Based on these trial searches, we ended up with a relatively simple search string: ''data market'' OR ''data markets'' OR ''data marketplace'' OR ''data marketplaces.''

Selecting the right scholarly search engines is an important step in structured literature reviews [41], [53]. In this review, only academic search engines were because several works focusing on grey literature and data markets in industry already exist [14], [15], [21]. Since the goal of this structured literature review is to consider as many domains our possible, our selection of search engines should cover as a wide a scope as possible. To this end, our selection was guided by other structured literature reviews [41], [54] and the platforms that were used in our search are *Google Scholar*,[9] *ACM Digital Library*,[10] *SCOPUS*,[11] *IEEE Xplore Digital Library*,[12] *Science Direct*,[13] *Wiley InterScience*,[14] *PiCarta & WorldCat*,[15] *JSTOR knowledge storage*[16] and, *ProQuest ABI/Inform*.[17] For each of these search engines we manually chose appropriate search options (e.g., to search only in title, abstract or keywords) which can yield more desirable results [55]. Furthermore, some search engines returned too many results to realistically process (e.g., Google Scholar returned approximately 142.000 results). In these cases we only considered the 100 most relevant results.

After querying each scholarly search engine, we narrowed down our selection by using several pre-defined exclusion criteria. Following this initial selection, a pre-defined quality criterion (see below) was applied to reduce the selection to a more manageable size. Finally, the abstract of each work was analysed manually to ensure relevance for answering the research questions. Below, these exclusion criteria are discussed; note that fig. 3 shows how many works were excluded during each of these steps.

[9] scholar.google.com
[10] dl.acm.org/
[11] scopus.com
[12] ieeexplore.ieee.org
[13] sciencedirect.com
[14] onlinelibrary.wiley.com
[15] Together, through our university's library: picarta.oclc.org, worldcat.org
[16] jstor.org
[17] about.proquest.com/products-services/abi_inform_global.html

16) EXCLUSION CRITERIA

As a first step, we remove duplicates that were found through multiple search engines. Moreover, we only consider works that were written in English. Finally, to narrow down our selection before the next step, we remove any papers that were written in 2015 or before. Since data markets are a recent phenomenon, this means we keep most of our papers while allowing us to focus on the state-of-the-art.

17) QUALITY- AND CONTENT-BASED EXCLUSION

As a way of guaranteeing the quality of the selected works, we filter on format as suggested by [42], [54]. We select only scholarly articles and books while excluding workshops, technical reports, and other formats encountered during our initial selection. After excluding undesirable formats, each remaining work's abstract was analysed manually to assess whether its content was relevant for answering our research questions [41], [53]. During this step, the selection was guided by four characteristics, directly derived from our research questions; a work was included for the next step if it:

- Proposes a data market as a solution to a problem or evaluates an existing data market,
- Proposes or evaluates a specific design, definition, formalisation or feature of a data market,
- Identifies challenges for implementing data markets or solutions for such challenges,
- Gives an overview of different data markets.

If a work appears to fit any of these criteria based on its title, abstract and keywords, it is included in our final selection.

Finally, since existing meta-research on data markets has already covered different types of pricing mechanisms [21]–[23], we excluded those works whose main contribution, based on their abstract, concerns pricing mechanisms.

B. EXECUTION PHASE

During the execution phase, the review protocol, as defined in the previous section, was applied. The execution lasted from March 2021 to June 2021; initially, 682 works were found, and this number was reduced to 82 based on the various selection criteria discussed above. We used the ''Mendeley Desktop'' tool[18] to manage the works while processing them. An essential feature of this tool is that it enabled the researchers to work on the same set of documents by sharing them via the cloud.

Figure 4 shows the distribution of the final selection over the years, and fig. 5 shows the distribution of the final selection over the different scholarly search engines. As can be seen, there is a definite increasing trend in our selection up until 2020. Based on fig. 1 we hypothesise that there was no decrease in publications in 2020, but instead that this is due

[18] https://www.mendeley.com/download-desktop-new/

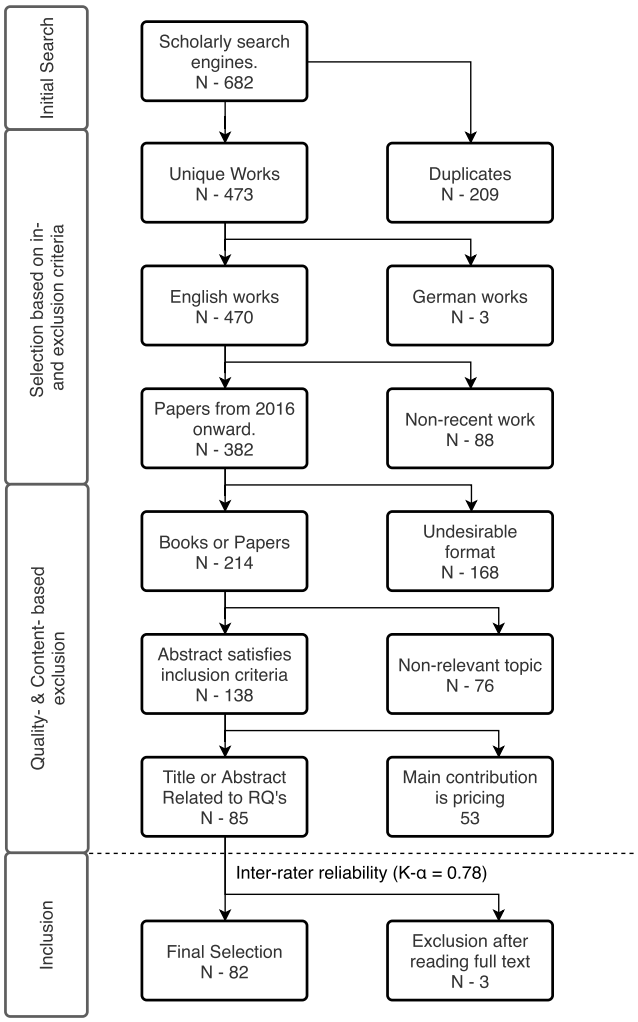


FIGURE 3. **The search and selection process. The steps are shown sequentially with the number of included and excluded works. After the final exclusion step, an inter-rater reliability test was performed and Krippendorff's alpha coefficient was calculated to ensure reviewer bias was not too high.**

to the fact that recent works have attracted fewer citations [56], [57] and speculate that this might make them less likely to be considered relevant by academic search engines and, consequently, to appear in our results. All search scholarly search engines contributed unique works, except for JSTOR knowledge storage which did not provide *any* works that made it to the final selection and is excluded in the figure.

In order to counteract potential subjectivity during the inclusion assessment of works based on their relevance, each abstract was read independently by two researchers who classified the work as relevant or not. Disagreements in classifying a work based on quality- & content-based exclusion were resolved through discussion by the two researchers who classified that work. In order to assess the inter-rater reliability of the inclusion assessment, the Krippendorff test was used to assess the agreement achieved among observers who categorise a given set of objects [58]. The Krippendorff

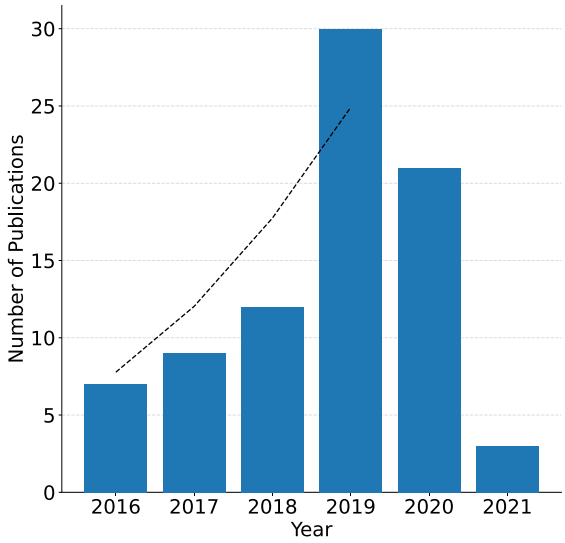


FIGURE 4. **Distribution of the years of publication for the final selection. An exponential growth in the number of publications is observed up to 2019.**

test yields an α-coefficient (K-α) which quantifies this agreement. Based on the inclusion- and exclusion criteria, a K-α of 0.78 was achieved, which indicates that the reviewers' bias was not too high.

C. REPORTING PHASE

In order to answer research questions 2 through 4 grounded theory was applied as a means to synthesise the data. Grounded theory is a systematic methodology for qualitative research with the ultimate goal of deriving theory from a large number of textual sources [45], [46]. Each text is read carefully, and useful concepts are identified by marking key sentences as *codes* that succinctly summarise them. Initially, this process yields many different codes, but as more texts are processed, they are merged and refined, leading to groupings of codes called *concepts*. Similar concepts can then be identified as *categories* which are, finally, used to derive a theory that provides insights into the subject of research [59] (see also table 2).

In the scope of the structural literature review, the coding process was used to derive the possible values for the five dimensions shown in table 1 that address research questions 2 through 4. Whenever a domain, role, definition, problem or solution was explicitly mentioned, it was marked. As more texts were processed, the concepts were merged and refined, leading ultimately to the values shown in table 1. Because the study selection strategy aims at gathering works from as many application domains as possible, the synthesis ultimately leads to a comprehensive taxonomy of the relevant problems and solutions.

IV. RESULTS

In this section, the results of the execution phase are presented; data from the extraction form is synthesised and

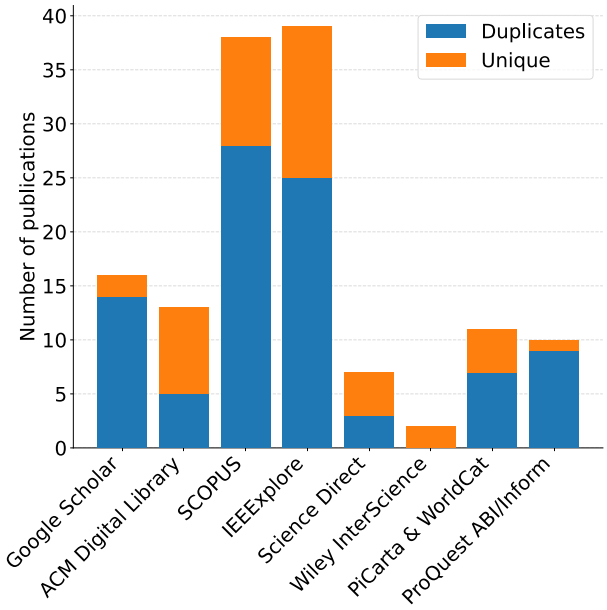


FIGURE 5. Distribution of the different scholarly search engines for the final selection. Most works are duplicates and most search engines contributed unique works to the final selections.

TABLE 2. The four levels of grounded theory and their scope.

Level	Scope
Code	Marking of individual passages related to research questions.
Concept	Groupings of codes that are conceptually coherent.
Categories	Collections of similar concepts that can be used to derive a theory.
Theory	An overview of the different categories that provides insight into the subject.

visualised to address the proposed research questions with one sub-section per research question.

A. RQ1: TYPES OF DATA MARKETS

In order to answer RQ 1, the organisational dimensions of data markets that have been identified and investigated in previous literature reviews for data markets in industry are considered. First, the statistics of each dimension are discussed and considered in the academic context of this work. Afterwards, whenever possible, a comparison is made with the results from the previous surveys on data markets in industry by Stahl *et al.* (2015) [14], and Spiekermann (2019) [15]. One important thing to note is that the values in the dimensions are often not mutually exclusive; for example, a literature review can consider *all* types of matching models and, consequently, the sum of the percentages of each value in a dimension can exceed 100%.

MATCHING MODEL

Figure 6 shows the distribution of the different matching models considered during the structured literature review. By far, the most discussed matching model is many-to-many matching, considering data markets in which data providers

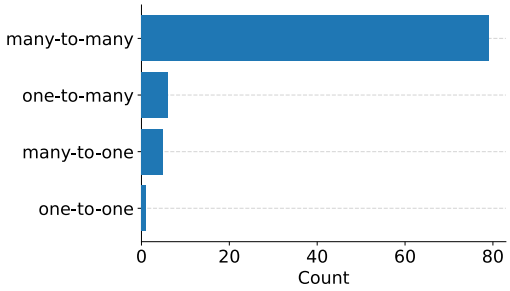


FIGURE 6. Matching Models in the structured literature review. Many-to-many is by far the most popular matching model.

provide data to multiple data consumers and data consumers consume data from multiple data providers. In contrast, one-to-one matching was only considered once: in the literature review of Stahl *et al.* which motivated our inclusion of the matching model dimension. Finally, the one-to-many and many-to-one matching models are mentioned five times and six times, respectively, usually together. The most common manifestation of these types of data markets is for companies whose business model already involves collecting lots of data (e.g., health wearable manufacturers [60] or music streaming websites [36]). These companies use many-to-one matching to act as private buyers for collecting data on the one hand and one-to-many matching for selling data as private sellers on the other hand. No previous survey explicitly considered the matching model dimension in their results; however, an investigation of the data markets investigated by the more recent work of Spiekermann shows that twelve out of the fourteen (≈86%) data markets investigated are designed with a many-to-many matching model. The other two platforms can be used to develop either internal (many-to-many) data markets for companies but could theoretically also be used for one-to-many data markets. Thus, it seems that the academic focus on many-to-many models is in line with the trends in industry.

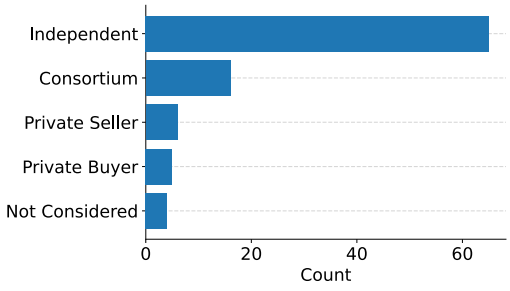


FIGURE 7. Platform Owners in the structured literature review. Private sellers and buyers are relatively rare.

PLATFORM OWNER

As can be seen in fig. 7, the results for the platform owners dimension are very much in line with the expectations set by the matching model dimension: Each time a many-to-one or one-to-many matching model is applied, the platform

is owned by a private buyer (5 instances, ≈6%) or private seller (6 instances, ≈7%), respectively [36], [60]. A notable exception is presented Halpin *et al.* [61] whose work sets out to facilitate crowd-sourcing of high-quality data. To achieve this, they propose that data consumers pay many ''experts'' to create task-specific data. In this way, the data market provides many-to-one for *multiple* data consumers, and the platform can be independently owned. For data markets with many-to-many type matching, we see 16 (≈20%) works where data markets are owned and operated by consortia [62], [63], and 65 (≈79%) where they are independently owned. Finally, it is interesting to note that 4 out of the 82 selected works (≈11%) do not consider the platform owner at all, despite the platform owner's clear implications on design choices for data markets [12], [64]–[66]. This distribution contrasts starkly with the survey of Stahl *et al.* of six years ago. Out of 72 data markets investigated there, 54 (75%) were privately owned, consortia owned 6 (≈8%), and 12 (≈17%) were independently owned. Again, Spiekermann does not provide any statistics themselves, but our own investigation reveals that only two out of fourteen (≈14%) data markets in his investigation can be considered privately owned, whereas the other twelve (≈16%) are independently owned. These results seem to indicate a shift away from privately-owned data markets towards consortium- or independently owned data markets that, for the most part, is in line with the focus in academic works.

PLATFORM ARCHITECTURE

Interestingly, fig. 8 demonstrates that most approaches favour a decentral approach (42 instances) over a central approach (23 instances). Eight works consider both types of architectures, most of which are literature reviews, but there are two exceptions: Spiekermann *et al.* describe a metadata model which they envision can be applied in both central and decentral data markets [67]. Additionally, Fernandez *et al.* propose a toolbox for designing data marketplaces with either a central or a decentralised architecture [34]. Disregarding the nine works that did not consider their architecture, this means that approximately 42% of all approaches consider a centralised- and approximately 68% of all approaches consider a decentralised architecture for data markets.

This trend is opposite to the observations of Spiekermann, who finds only four out of fourteen (≈29%) data markets that follow a decentral architecture versus eight (≈71%) that use a central architecture. One possible explanation for this gap between academics and industry is the surge in academic interest in decentralised software systems such as blockchain technology which is not yet finding much counterplay in practice [54].

TIME FRAME

Figure 9 shows that data markets are usually being investigated for either dynamic data (37 instances) or static data (28 instances), and only 16 instances consider data markets that trade in both types of data. Ignoring the work that did not consider a time frame for the data, this means that approximately 65% of academic works consider dynamic, and approximately 44% of academic consider static data. When comparing this to Stahl *et al.*'s survey results from 6 years ago, a trend is observed towards more dynamic data; in his results, approximately 40% of the data markets offer dynamic data, and approximately 86% of data markets offer static data.

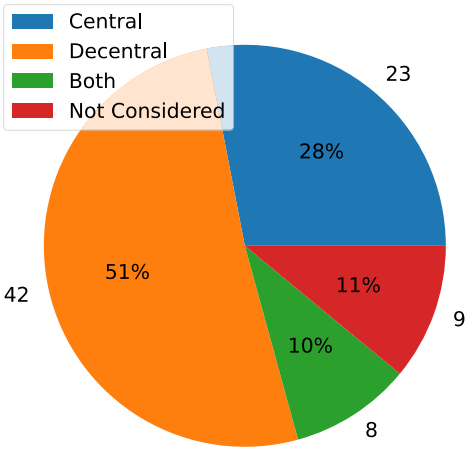


FIGURE 8. Platform Architectures in the structured literature review. More than half of all works consider a decentral architecture.

DATA ACCESS METHOD

Figure 10 shows the distributions of all data access methods for the data markets covered in the structured literature review (in grey). The most commonly considered data access method is direct download (31 instances), closely followed by push-based and pull-based access (25 instances each). Specialised software and compute-to-data approaches are less common (12 instances and 10 instances, respectively), and 7 works did not explicitly consider any data access method at all. Based on our results, we identify two types of specialised software: centralised platforms for data manipulation and decentralised distributed ledger-based applications. An example of the first variety is the FIWARE platform which is a European Union reference platform for IoT [68] and which is used by the BONSEYES Data Market [62]. On the other side of the spectrum, the IOTA protocol is often used [7], [69]–[71]. IOTA differs from traditional, blockchain-based, decentral, distributed ledgers because it employs a non-cyclic graph for securely sharing data [72].

It is clear that some data access methods are designed for specific time frames: e.g., direct download and compute-to-data make more sense for static data than dynamic data, whereas push-based methods such as streaming are designed with dynamic data in mind; thus, we expect the data access methods to vary significantly based on the time frame of the data products. In order to verify this hypothesis and identify

other potential patterns fig. 10 also shows how often each data access method occurs for both static (in orange) and dynamic (in blue) data products. In order to ensure that the data access method and data product belong to the same data market, we excluded literature reviews for fig. 10. As can be seen, on the one hand, there is a clear trend where direct download and compute-to-data are more popular for static data (20 instances and 6 instances respectively) versus dynamic data (5 instances and 3 instances, respectively). On the other hand, push-based and pull-based access methods are used significantly more for dynamic data (20 instances and 16 instances, respectively) than static data (4 and 5 instances, respectively).

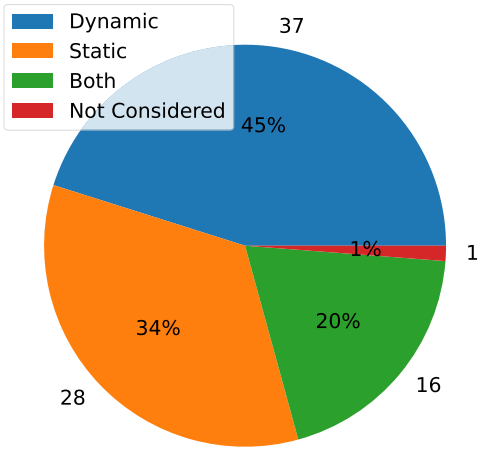


FIGURE 9. Time Frame of data products in the structured literature review. Markets focusing on dynamic data are most common, followed by markets for static data and, finally, markets that allow for both types.

It is not easy to compare these results with the findings of Stahl *et al.* accurately six years ago because, as was noted in section III-A the values they used do not apply very well to current academic research. Nevertheless, it can be observed that downloading and push- and pull-based APIs were the most popular approaches in industry in 2016 and still are in academia in 2021.

ARTEFACT

The final dimension considers the artefact investigated in each of the investigated works. As can be seen from fig. 11, there is a good amount of both full data market architectures (38 instances) and methodologies for data markets (20 instances), as well as a significant number of literature reviews (14). At the same time, our results indicate that most academic work currently lacks an evaluation in an industrial setting: only seven instances of data market architectures, nine methodologies and one literature review were conducted with an industrial partner. On the one hand, this lack of real-world evaluation could be explained by one of the main conclusions of both the surveys by Stahl *et al.* and

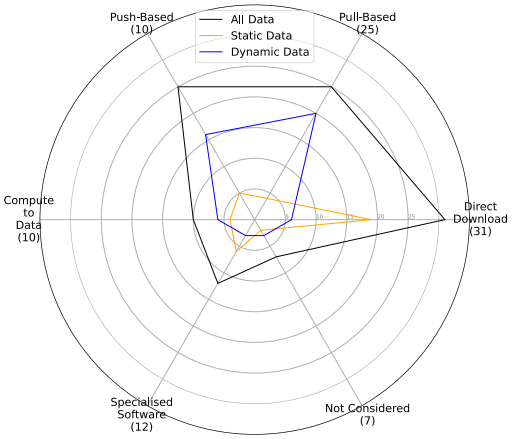


FIGURE 10. Data Access Methods in the structured literature review. Specialised approaches, including compute-to-data are less common.

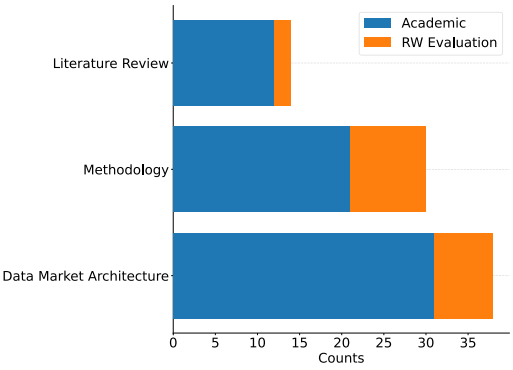


FIGURE 11. Artefacts investigated in the structured literature review. Most works do not evaluate their artefact in the real world.

Spiekermann: many, if not most data markets in industry fail to be profitable and close after a few years.[19] On the other hand, the lack of real-world evaluations could be the reason for the discrepancies between academic work on data markets and industrial surveys.

B. RQ2: DATA MARKET CONTEXT

Artefacts and data markets, and methodologies for them in particular, do not operate in isolation. Design science tells us that it is imperative to consider the context that is relevant for and the roles of the actors that interact with the artefact under design [43], [44]. In this section, the application domains of data markets in the academic literature are considered and explained as well as the roles that have been identified in each of these domains. In doing so, not only is RQ 2 addressed but also the way is paved for answering RQ 3 and RQ 4. After all, problems exist within a *problem context* and are experienced by *actors taking on a role* who interact with the artefact. Similarly, solutions are applied only within a context and attempt to meet the requirements set by the same roles

[19]The most prominent example of a data market closing is perhaps Microsoft's Azure data market [73]

that experience problems. Finally, different definitions of data markets were collected and examined. These definitions ultimately led to the minimal definition of a data market presented in section II.

APPLICATION DOMAINS

The application domain is the domain in which the data market(s) discussed in each work operate and from which the researchers that propose the data markets themselves hail. A single work might consider multiple application domains; for example, a literature review could consider both data markets for hospitals and health institutions as well as data markets that trade user behaviour gathered in online environments [34]. Moreover, the domains identified through grounded theory cannot always be viewed as entirely separate: healthcare data is generally also personal data, and machine learning can be done on industrial data. Nevertheless, the identified domains, which are shown in fig. 12 are more or less cohesive concepts, each with their own emphasis on specific roles, problems and solutions.

A striking finding of the structured literature review is that a little over half of the works identified (46 works) involve data in the context of Internet-of-Things (IoT). Internet of Things is an approach whereby a network of physical objects (things) are embedded with sensors and connected to the internet to exchange data [74]. Since the facilitation of exchanging data is the primary concern of data markets, it makes sense that data markets are often investigated in this context. Furthermore, the IoT approach is itself applicable to many application domains [75], which is also reflected in our results, although some domains lend themselves more readily to IoT than others. Moreover, not every domain where the IoT approach is applied produces IoT data exclusively and in most domains, both types can be observed. To better investigate the different domains and the prevalence of IoT in data markets, fig. 12 shows (in orange) for each domain the number of data markets identified that cater towards IoT data in that domain.

Below, these domains are discussed, starting with the domain-agnostic and Personal Data domains. Afterwards, the domains in which IoT data markets are more prevalent are discussed and, finally, the domains in which IoT is rarer.

1) DOMAIN-AGNOSTIC DATA MARKETS

Interestingly, only about a quarter of all cases (24 works) do not explicitly consider a specific domain for the data that is traded on them; we consider these to be *domain agnostic* data markets. Half of the literature reviews in the study (7 works) are domain agnostic, meaning that only 15 works with just one (type of) data market are domain agnostic. The clear benefit of domain agnostic data markets is that, in theory, they can trade data from all domains, which in turn means that it is possible to attract a large number of both data providers and data consumers. However, the other side of this is that it is hard to consider all the context-specific considerations such as the roles, their problems, and the corresponding solutions necessary to trade data regardless of their domain. Indeed, upon closer inspection, we find that (excluding literature reviews) most data markets (12 out of 15, 80%) only facilitate the exchange of static data, which is a straightforward way to address this complexity.

Contrary to domain agnostic data markets, most academic literature focuses on trading data within a specific domain. This focus limits the number of data providers and data consumers that can be attracted but has the advantage that the context and roles are more clearly defined when designing artefacts. This observation validates one of the main contributions of this literature review: a structural discussion of the different application domains, roles, and the problems and solutions that originate from these. Besides general IoT and Domain Agnostic, the review yielded 8 application domains with varying frequency, ranging from 12 instances of personal data markets to 2 instances each of smart home- and government data markets.

2) PERSONAL DATA

The personal data domain is divided relatively evenly between IoT (7 instances) and non-IoT data markets (5 instances). The domain follows a definite industry trend whereby new business models are arising to monetise personal data [16] or capture the value of personalised services based on personal data. Personal data comes in many shapes, including data collected through interaction with online platforms [36], credit scoring information [76], and personal IoT sensor information [20]. An important reason for the prevalence of research in the personal data domain is the intersection of many interests and domains. Besides the purely economic perspective, which is typical for all domains, there are explicit ethical [16] and legal [77] dimensions to be considered. These dimensions are driven by the high risk of abuse of personal data [71], as well as attempts by multiple governments to regulate the collection and processing of personal data. Examples of the latter variety include the General Data Protection Regulation (GDPR) in the European Union [78] and the California Consumer Privacy Act (CCPA) in the state of California [79].

Next, we discuss those domains where IoT data markets are prevalent; these include 'automotive & mobility', 'smart cities', 'healthcare', and 'smart homes'.

3) AUTOMOTIVE & MOBILITY

Automotive & mobility data markets concern data that is generated by vehicles, which are both the consumers of data services and the providers of data [80]–[83] but also public and private transportation providers [84]. The automotive and mobility domain is characterised by many independent actors (e.g., travellers, car manufacturers and public transport providers), whose collective behaviour is of interest for providing services (such as traffic prediction or resource scheduling). Consequently, the data market platforms are

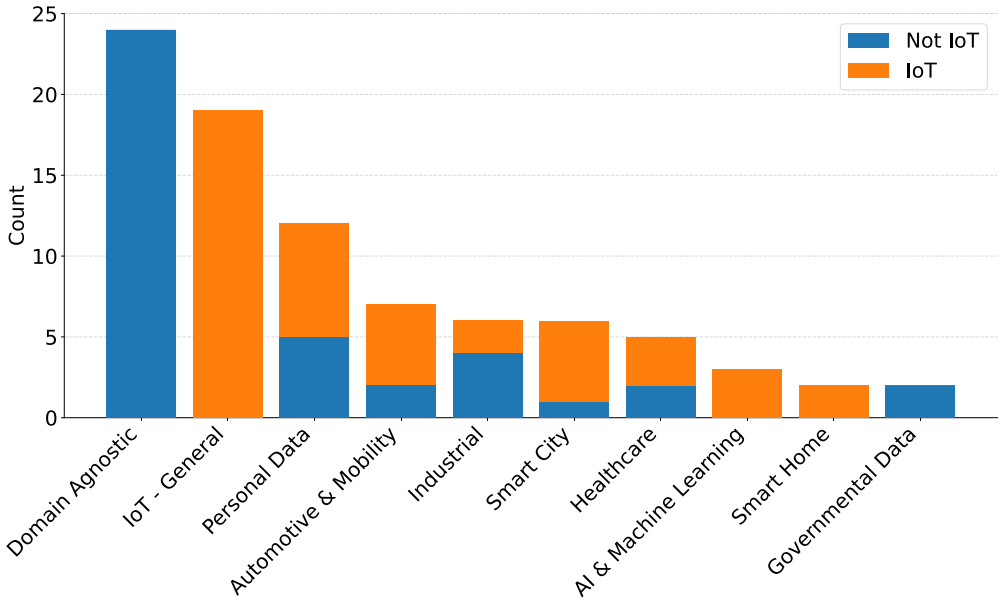


FIGURE 12. Application Domains. About half of all works on data markets consider IoT data of some kind and in general IoT data is more common in some domains than it is in others.

always owned independently or by consortia. Finally, given the fact that modern-day vehicles are often outfitted with sensors [85] that become the ''Things'' in IoT, it makes sense that these data markets are investigated from an IoT, dynamic data perspective.

4) SMART CITY

Another domain where IoT data markets are dominant is the smart city domain. Many definitions of the term 'smart city' have been proposed [86], but a good informal definition is that *"a smart city enhances citizen life quality by solving city problems"* [87]. This achievement, in turn, is generally achieved by the provision of services driven by data collected by (IoT) sensors. Data markets facilitate data exchange between the data consumers, which provide the services, and data providers, which own or operate the sensors. One peculiarity of data markets for smart cities is that, as the foremost representative of a city and its citizens, local governments will always play an outsized role as a data provider, and often also as a data consumer [17], [88].

5) AI & ML

Three works focus on data markets for Artificial Intelligence (AI) and Machine Learning (ML) [50], [52], [62]. All of these are concerned with IoT data and are aiming to facilitate so-called *edge computing*. In the edge computing paradigm, computation happens close to the data instead of in a centralised (often cloud-based) environment, which is expected to improve performance [89]. Data markets designed for these purposes envision that the IoT device owners make available an environment to facilitate a compute-to-data access method.

6) HEALTHCARE

Healthcare data might be considered specific to a particular industry, which would qualify it as industrial data, but it is set apart by an important distinction: the data providers are generally individuals (patients) rather than companies. Traditionally, healthcare data has been static (e.g., medical or DNA records [8], [90]), but it can also come from IoT sensors, such as wearable devices [20], [29] or smart homes [64] in which case the traded data is dynamic. Healthcare data can almost always be considered personal data, especially when the data provider is an individual and the data is not aggregated. Consequently, most of the work on personal data markets is also relevant to healthcare data markets.

7) SMART HOME

Finally, two works focus on data markets for smart homes. Smart homes are IoT applications whereby sensors are placed in a home to measure and control properties such as temperature, humidity, safety, energy consumption and sometimes even health data [64], [91]. Third parties who desire to investigate behavioural and consumption patterns can also use the data from each of these homes. Data markets facilitate the option for smart home owners to act as data providers to exchange their data to these third-party data consumers. The data can be exchanged for money, but it is also common for the smart home owners to be rewarded by increased functionality or improved services from the devices that make up the smart home.

Next, we discuss the domains where IoT data markets are less common: industrial and governmental data.

8) INDUSTRIAL

The industrial data domain focuses on data that is specific to particular industries. For example, several works are intended for data-driven manufacturing in the context of the 'Industry 4.0' initiative [92]. In this context, data is traded both *internally* [33] as well as between manufacturing companies [93], and these companies are both the data providers and the data consumers. Internal trading can happen between departments that are organisationally or geographically separated and that want to share their data, similar to a *data mesh* whereby data products are offered by different domains [94]. External trading happens in industries where cooperation between different companies or organisations is necessary. Examples for which data markets are being investigated include recycling [95], energy [96], and more generally supply chains [9], [31].

9) GOVERNMENTAL

Finally, the governmental data domain was identified based on only two works and is characterised by data providers who are one or more government agencies. Contrary to data markets in the industrial domain, governments are mostly interested in disseminating their data through which they hope to inform their constituents or enable services that improve their lives. Consequently, data is generally publicly available for any data consumer who wants it, and the data market platform is oriënted towards creating a central infrastructure for exchange [17], [35].

ROLES

The data provider, data consumer and platform owner were already introduced as roles with the definition of a data market presented in section I and are present in every manifestation of a data market. These do not need to be discussed again, and, as such, they are left out of this section. As for the other roles, which will be discussed in this section, the terms used here are not universally accepted. In fact, many different terms are used sometimes with contradicting definitions. For example, data brokers have been defined as the actors responsible for modifying or processing data products before consumption [4] (i.e., a data transformer) instead of the actors responsible for acting as the intermediary that helps data consumers and data providers find each other. Moreover, it is common for one actor to take on multiple roles: In many centralised data markets, the data broker is also the clearinghouse as well as the infrastructure provider.

Nevertheless, each role discussed here can also be viewed separately and has been filled by independent actors in at least one work. Figure 13 shows the frequency of each of the roles as they are discussed in the works in the literature review. It can be observed that all roles are considered with more or less the same relative frequency for both IoT- and non-IoT domains, meaning that they are worthy of consideration for both types of data markets. Below, these roles are discussed in descending order of the frequency of their appearance in the literature review.

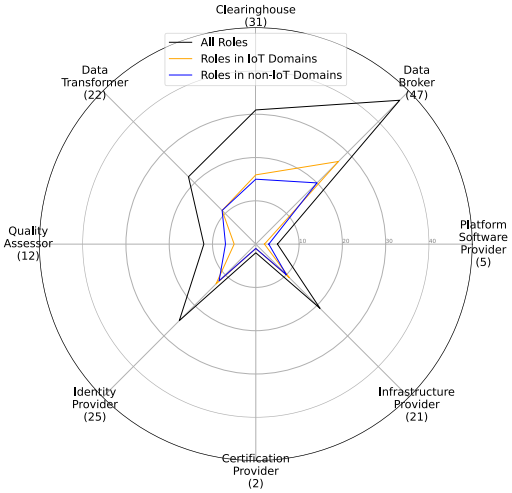


FIGURE 13. Roles found in the structured literature review. Both IoT and non-IoT data markets consider the same roles with more or less the same frequency.

10) DATA BROKER

The data broker is the actor responsible for acting as an intermediary between the data provider and the data consumer and matches their needs by helping the former present their data product(s) and the latter find suitable data products. Generally, this is a passive role that involves storing, managing and making available information about the data products on the data market [32]. This service is critical to the facilitation of data product exchange, which explains why it is the role that is most commonly explicitly discussed. In its most diminutive form, a data broker is just the actor who provides a registry of all data products (e.g., [97],) but there are more active manifestations as well, e.g., data consumers can take on the role of data broker when they formulate their requirements for a data product. In such a case, the exchange of data products can be negotiated in several rounds [34] or even purposefully created based on requirements [61].

11) CLEARINGHOUSE

The clearinghouse is a common intermediary in many types of (financial) markets. It is responsible for validating and finalising the transaction, ensuring that both the provider and the consumer honour the agreement set up through the data broker [98]. In the context of data markets, enforcing the transaction includes ensuring the ensuring that the data consumer gains access to the data product while at the same time ensuring that the data provider gets compensated. Who fills the role of the clearinghouse is highly dependent on the data market architecture: in a centralised data market, one clearinghouse (usually the same actor as the data broker) can enforce all transactions [96], but in a decentralised data

market, the role of clearinghouse might fall onto the data provider or could even be performed by one or more third parties that specialise in transaction enforcement [90].

12) IDENTITY PROVIDER

Since data trading occurs mainly through (online) IT platforms, it can be challenging for the different actors to establish with whom they are dealing. Therefore, implementing some identity tracking, even an anonymised one, is also necessary for the data broker and clearinghouse to operate normally. Identity providers offer a service to provide identity information to different actors. This service can be as simple as assigning actors a digital object identifier (DOI) [70] but can also involve keeping extensive profiles on different actors with real-world identity information [32].

13) CERTIFICATION PROVIDER

The certification provider goes one step beyond the identity provider by providing certificates to different actors based on verification or evaluation. The certificates can be used simply as an indication of trustworthiness or quality [96], they can signal compliance with specific quality standards [75] or even be used to keep track of permissions for different actors [32].

14) DATA TRANSFORMER

When designing data markets, it is often envisioned that data products that have been made available through the platform need to be processed further before they can be usefully consumed. This transformation is the role of the data transformer, who is one of the most versatile actors in a data market. Data transformation can encompass everything from cleaning, aggregating, standardising or creating new data products from existing data [34], [35]. In their most basic form, data transformers combine the role of data consumer and data provider: consuming data, transforming it by themselves and then providing the transformed data as a new data provider [99]. Nevertheless, data transformers are often also separate actors when data markets are designed with data transformation as a specific requirement or build-in feature [28].

15) INFRASTRUCTURE PROVIDER

Processes in a data market, such as data brokering and transaction enforcement, require the deployment of an extensive IT infrastructure. The infrastructure provider is the actor responsible for providing this infrastructure. In centralised data markets, the infrastructure provider is often the platform owner. However, it is also common to have independent, third-party infrastructure providers, either when the data market provider lacks the technical skills or when data market architecture is decentral [32], [84], [100].

16) QUALITY ASSESSOR

The quality assessor is a role for actors who assess the quality of a data product according to some guidelines. This assessment aims to make it easier for data consumers to find and select useful (high-quality) data products and motivate data providers to consider the guidelines for providing high-quality data products. The quality assessor is often automated by using standardised tests [33], but some solutions develop methods specifically to leverage manual quality assessors [61], [101].

TABLE 3. Problems that are not specific to data markets. In total 63 works considered one or more of these problems, most commonly Security.

Problem Category	Frequency
Security	53
Scalability	16
Performance	14
Strategic Behaviour	9
Efficiency	3
Resource Minimisation	2
Ethical Concerns	1
Maintainability	1
User Friendliness	1

17) PLATFORM SOFTWARE PROVIDER

As data markets evolve, the platform software provider is the role taken on by the actors responsible for designing and maintaining the software necessary to run the data market. As fig. 13 suggests, this role is not often considered explicitly in academic literature, but their role is important nevertheless. Providing platform software is especially challenging in a decentralised environment, where different actors have to achieve consensus on the software they use to facilitate data exchange [83].

C. RQ3: PROBLEMS FOR DATA MARKET DESIGNS

Now that the context and roles for data markets have been established, we turn towards addressing RQ 3 by discussing the different challenges for designing data markets that arise from the application of grounded theory in the literature review. It is important to note that many of the problems identified in this way come down to achieving well-known software architecture quality attributes [102] and are not specific to designing data markets. Since this literature review focuses on data markets, problems common for designing any data-intensive IT artefacts are excluded from the resulting taxonomy. These problems include achieving efficiency, ethical concerns, maintainability, performance, reliability, resource minimisation, scalability, user-friendliness and strategic behaviour by malicious actors, and the frequency of their occurrence in the literature review is shown in table 3. Achieving these (extra-functional) quality attributes is generally the problem of the system architect, platform software provider and platform & infrastructure providers.

Figure 14 visualises the taxonomy of the identified problem categories that are specific to data markets, based on the results of the literature review. It can be read as follows: the outer two columns show the main roles from section IV-B who are tasked with addressing the identified problems, while the central column shows whether the problem primarily

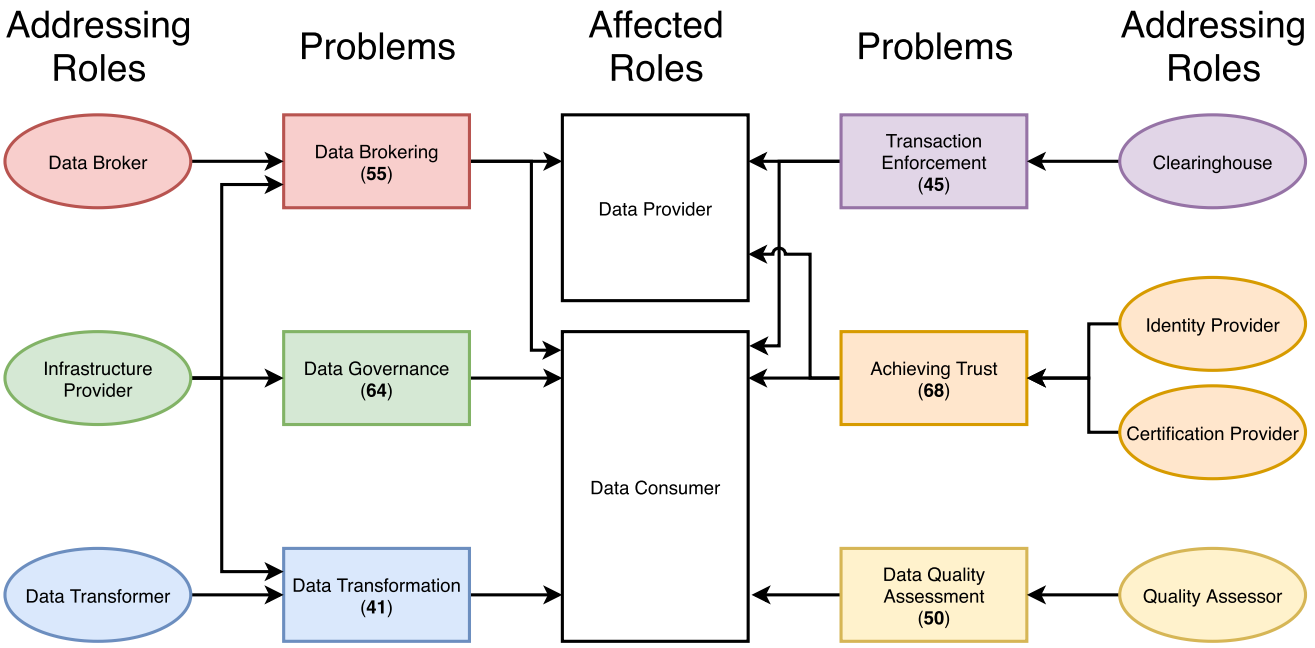


FIGURE 14. Taxonomy the problems in data market design. The outer two columns show the main roles from section IV-B who are tasked with addressing the identified problems, while the central column shows whether the problem primarily affects the data provider, data consumer, or both. The problem categories are shown in the second and fourth columns; the frequency of their occurrence in the literature review is also shown in bold font for each problem category. Problems and their most important corresponding addressing roles are shown in the same colour to group them together.

affects the data provider, data consumer, or both. The problem categories are shown in the second and fourth columns; the frequency of its occurrence in the literature review is also shown in bold font for each problem category. Problems and their corresponding addressing roles are shown in the same colour to group them together. The infrastructure provider is an exception, as it addresses more than just the data governance problem category but is coloured green because there are no other main roles that address these problems (other than the platform software provider and the data consumer).

Three important caveats have to be noted when considering fig. 14: first, the platform software developer role is missing in the figure. This is because all the problem categories rely on the platform software, and therefore the platform software developer (indirectly) addresses all the problem categories. Secondly, the affected roles (the data provider and the data consumer) are also partly responsible for addressing the problems that affect them. Finally, data valuation efforts (such as how to price data) are excluded from the results as the methodology explicitly excluded papers whose main contribution was pricing mechanisms.

Below, the problems in fig. 14 are discussed, as well as how they affect the data provider and data consumer.

1) ACHIEVING TRUST

Despite its inclusion in fig. 14, achieving trust in the data market platform is not a problem specific to the domain of data markets. Nevertheless, there are two reasons for its inclusion. Firstly, including achieving trust in fig. 14 allows us to include the identity providers and certification providers whose main contribution to the data market platform is to help achieve trust. Secondly, as is argued below, trust is an essential *solution* to dealing with uncertainty [103], [104]. Since trust is actively investigated in the context of data markets, fig. 14 includes achieving it as a problem.

2) DATA BROKERING

When designing a data market, it is essential to consider how the data consumers are matched with data providers. This problem, called data brokering, shows similarities with querying challenges in database management [105] and search engine design [106], but some distinct differences complicate data brokering in data markets. Firstly, data providers and data consumers are often anonymous, and there can be no assumption of mutual trust [107]. Moreover, the novelty of data markets and the corresponding lack of standards make it difficult to express in a structured manner both what data product the data provider is offering, as well as the desired properties of the data product that the data consumer is looking for [34]. Indeed, because data is an experience good [13], its usefulness is highly dependent on the context of the data consumer; e.g., what other data need to be integrated with it. Finally, we note that the data consumer is not the only role with requirements as the data provider can also expect to be somehow rewarded for the exchange. As was previously noted in the introduction, the price determination problem has already been extensively covered in other works [22], [23]. However, other mechanisms for rewarding data providers can

also be considered, such as services [64], [91] or company rewards for internal data markets [33], [94].

3) DATA GOVERNANCE

Data governance is an oft-used term in the context of data management, where it pertains to the ability of an organisation to ensure the existence of, and control over, high-quality data [108]. In the context of data markets, the problem of maintaining data governance can be expressed as defining and verifying the requirements for exchanging or distributing data products [109]. Simply put, data governance is the problem of maintaining ownership, and control over one's data before, during, and sometimes after an exchange.

An important subcategory of data governance problems is maintaining data sovereignty. Data sovereignty concerns the idea that data are subject to the laws of the country where they are collected [110]. The challenge with respecting data sovereignty is, on the one hand, finding a balance between protecting the data provider's data against illegal use and usefully sharing these data with a data consumer on the other hand [111]. In academic literature, the most commonly discussed aspect of data sovereignty is the notion of privacy, which can be defined as *'the claim of individuals, groups or institutions to determine for themselves when, how, and to what extent information about them is communicated to others'* [112]. However, other aspects, such as copyright protection or intellectual property laws, have to be taken into consideration as well [13]. Finally, it should be noted that metadata is also information that can be subject to laws. Accordingly, when designing a data market, all metadata, such as the geographical location of the data, the data providers and the data consumers, needs to be considered, and methods need to be integrated to enable data providers and data consumers to enforce and demonstrate compliance respectively.

Although data governance and sovereignty are closely related, data governance is the strictly greater problem category. Data sovereignty is concerned with requirements on data management imposed by *legal* requirements, whereas the problem of maintaining data governance concerns the requirements imposed by the *data provider*. As an example, the data provider can choose to remove their restrictions in exchange for additional compensation from the data consumer [113], [114]. However, this approach to data governance cannot be used when restrictions are imposed to maintain data sovereignty [16].

A significant part of the problem of data governance results from the low cost of replicating data [13]: once a data product has been consumed, it is easy for the data consumer to replicate the data and exchange it with other parties, effectively leaving the original data provider without control over their asset [9]. In its simplest form, the problem of maintaining data governance can be achieved by simply imposing restrictions on which actors can access what data [115]. Often, however, this is not enough, and the way data is consumed also needs to be monitored [34]. In such cases, managing data in a compliant manner requires specialised software to keep data from being abused. Designing such software is especially challenging in data markets, where no *one* actor is in charge of the data. In such data markets, third-party infrastructure providers can provide a specific infrastructure or environment that ensures data governance.

4) DATA TRANSFORMATION

Sometimes data products on a data market need to be processed further before they can be usefully consumed. Therefore, when designing a data market, one has to consider the problems of *data transformation*, such as how to aggregate, integrate, standardise or even add value to existing data products [35]. The manifestations of data transformation problems differ radically depending on whether the data market architecture is designed centrally or decentrally. In a central data market architecture, the platform provider can decide which transformations are desirable or necessary and arrange either for data providers to transform the data themselves or use third-party data transformers. On the other hand, in a decentralised data market, desirable data transformation must be inferred from or agreed upon by the different nodes and participants in the data market.

5) TRANSACTION ENFORCEMENT

When the brokering process is complete, and the data provider and data consumer have agreed on the terms for the exchange of the data product, the next challenge is to complete the transaction in a manner that ensures compliance with the agreement [116]. Ensuring a compliant exchange is the problem of *transaction enforcement*, and it can involve more than simply making sure that the data product is sent from the provider to the consumer (e.g., by providing middleware infrastructure or by having the client send it directly in a peer-to-peer fashion). For one thing, following the definition of a data market, the data provider should also stand to benefit from the transaction. Moreover, there is the more philosophical question of *ownership* and the transferring thereof.

Transaction enforcement is primarily the challenge of the clearinghouse and is highly dependent on three dimensions from RQ 1. First, the access method is quite obviously relevant for transaction enforcement, and indeed, specialised software is usually designed specifically to address the problem of transaction enforcement [72]. Second, the time frame dimension is connected to the access method and is of particular interest for transaction enforcement as dynamic data is generally consumed on a regular interval, or access is given over an extended period leading to multiple transactions instead of just one [33]. Finally, the data market architecture is a relevant dimension for transaction enforcement as it impacts where the responsibilities for enforcing transactions lie. In a centralised data market, the platform provider is generally tasked with enforcing transactions, whereas in a decentralised data market, the data provider and data consumer might have to transact in a peer-to-peer manner.

6) DATA QUALITY ASSESSMENT

In order to facilitate effective data trading, a data market should offer some method to assess or guarantee *data quality*. If no such method exists, the data market runs the risk of being flooded with low-quality data products, which in turn would discourage data consumers from using the platform [117]. Based on the works studied in this literature review, three facets of data quality have been identified: First, the quality of a data product has a service-level facet, which consists of extra-functional properties usually described in a service-level agreement (SLA), such as the availability of the data product and the responsibilities of the data provider and the data consumer [49]. Additionally, data quality has a content-based facet driven by the information contained in the data itself. This facet entails properties such as feature-richness (i.e., how many dimensions are in the data) [101] and truthfulness (i.e., the extent to which the data corresponds with reality) [118]. Finally, data quality has a context-based facet [119], which includes the ease with which it can be integrated with other data [34], the relevance of the data for the data consumer, as well as (proof of) data provenance (i.e., the origin of the data) [10], [13].

When creating quality metrics, the three facets of data quality identified above need to be considered. For example, the FAIR guiding principles stipulate that a data product should be Findable, Accessible, Interoperable and Reusable and specify how to achieve and measure these requirements [120]. The FAIR guiding principles address all three quality facets: metrics for findability and accessibility address the service-level facet, metrics for interoperability address the context-based facet, and metrics for reusability address the content-based facet.

D. RQ4: SOLUTIONS FOR DATA MARKET DESIGN

In this section, we discuss the different approaches that are being investigated to solve the problems identified above in section IV-C. Similarly to the problems, these approaches were identified through the application of grounded theory, whereby different solutions were captured, compared and aggregated until different codes arose, finally resulting in the taxonomy shown in fig. 15. In addition to the approaches (shown in rounded rectangles), fig. 15 shows how frequent each approach occurred in bold text. Furthermore, to improve clarity, each approach has been grouped and colour-coded by the problems (shown in squared rectangles) they solve; the arrows indicate which problems can be solved by which solutions.

It is important to note that, although most solutions individually address only one problem, some solutions become more effective when combined with others, e.g., a specialised querying mechanism can make use of anonymisation techniques such as differential privacy to address both the problem of data brokering and that of maintaining data governance. Other approaches naturally address more than one problem on their own; these approaches are coloured white and are connected to multiple problems in fig. 15. Finally, some solutions are coloured orange and connected to the squared rectangle called "Trust." Although there have been some works that consider achieving trust to be a problem for data markets in-and-of-itself [6], [18], [70], we argue that trust is, in fact, a solution for dealing with uncertainty, which is in line with previous work on the nature of trust [103], [104]. In this capacity, trust can simultaneously help solve all identified problems, especially when combined with other solutions for each problem. If all actors *trust* that the implemented approaches work as intended and solve their respective problems, then they will be more content, and, consequently, the data market will be perceived as more effective. The rest of this section explains the approaches shown in fig. 15 more detail; the implementation of each approach is briefly explained, as well as which roles are relevant to it and how it solves the corresponding problem(s) for these roles.

1) SMART CONTRACT-BASED AUTONOMOUS ACTORS

Blockchain and distributed ledger technology (DLT) have quickly become popular as means to manage data in a *distributed* and *decentralised* manner. These technologies aim to facilitate a database that is maintained by multiple, generally independent, actors who are guided by a *consensus protocol* to achieve consensus on the information that is in the database and how it can be updated [54]. A particularly interesting use case of blockchain and DLT is the *smart contract*, whereby code is stored as information on the blockchain and can be interacted with through transactions which are also stored on the decentralised, distributed database [54]. Smart contracts have been extensively researched in the context of data markets as means to improve trust in the ecosystem because [101]:

1) The distributed nature of blockchain and DLT means that the functionality of- and interaction with smart contracts is transparent (i.e., visible to all participants of the blockchain or DLT).
2) The decentralised nature of blockchain and DLT means that the execution of smart contracts is dictated by the consensus protocol and is not controlled by any individual actor, making smart contracts effectively autonomous after being deployed.
3) As with any piece of code that facilitates automation, autonomous actors improve the scalability of the ecosystem as they can process large amounts of data products automatically.

Despite these benefits, there are also several downsides to using smart contract-based autonomous actors: Firstly, smart contracts are particularly prone to abuse due to security threats, as their functionality is transparent and there is a lack of tools to assess smart contract security [121]. Moreover, a pitfall of automation is that the automated system cannot always handle corner cases that require special treatment [122]. In this literature review, six different

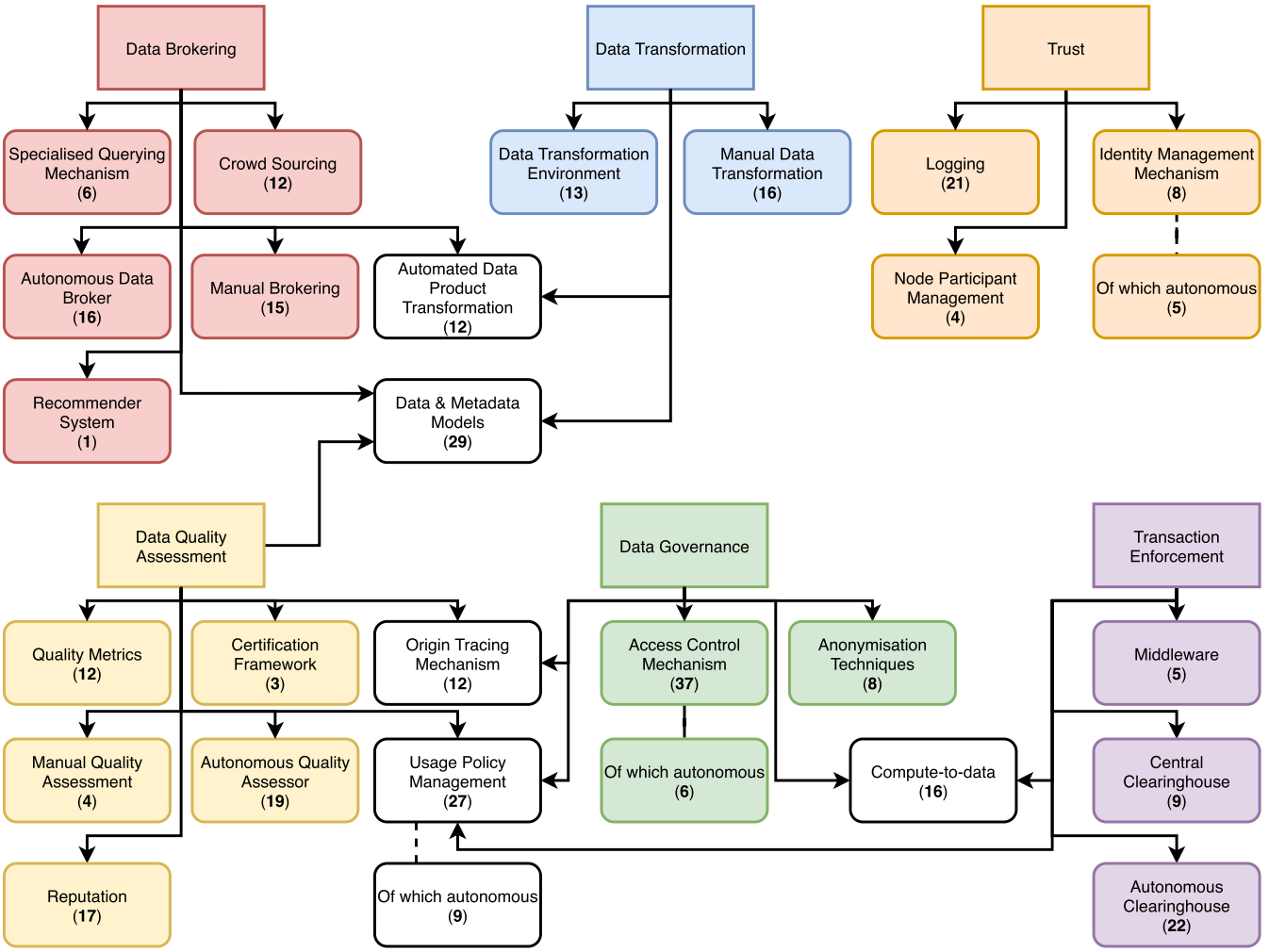


FIGURE 15. Taxonomy of the approaches for solving problems in data market design. Solutions are grouped and colour coded together with the problem they address with arrows indicating that a solution addresses a problem. White solutions address multiple problems, and the orange solutions grouped around "Trust" improve the effectiveness of all other solutions.

types of autonomous actors are identified, and 36 different works made use of one or more of these actors. The cases where the autonomous actor takes one of the roles identified in section IV-B are discussed first. Then, the autonomous usage policy management, access control mechanism, and identity mechanism actors are all discussed together with their respective non-autonomous implementation.

2) AUTONOMOUS DATA BROKER

As mentioned above, in its simplest form, a data broker is simply a registry where data providers can register their data products for data consumers to browse. The processes of registration and browsing can be easily automated and encoded in a smart contract [95], [123], [124], eliminating the need for a centralised registry. This approach has the advantage that it eliminates potential strategic behaviour on the part of the data broker, such as making certain data assets more easily findable [116].

3) AUTONOMOUS CLEARINGHOUSE

Another application for smart contract-based autonomous actors is the creation of an autonomous clearinghouse. A so-called *escrow contract* can be used in which a third-party escrow agent receives, holds and disburses assets according to a predefined contract or agreement on behalf of the agreement participants [125]. This approach can be adapted to smart contracts, where an autonomous clearinghouse receives and holds the payment for a data asset or a token for granting asset to a data product and makes sure these are exchanged according to the agreement between the data provider and data consumer [70], [115], [126]. In some data markets, the data itself is even stored on a blockchain in an encrypted manner [90], [127].

4) AUTONOMOUS QUALITY ASSESSOR

Autonomous quality assessment makes use of the fact that most consensus protocols make it practically impossible

to change historical data stored in the decentralised database [54] to store information related to the quality of consumption of data products. As discussed above in section IV-C, data quality aspects involve several facets, each with several aspects. Aspects from the service-level facet can be readily monitored and logged on the blockchain so that other data consumers can see the historical performance of the data product [70], [128]. The content- and context-based facets can be stored on the blockchain by asking previous data consumers to write reviews, which are then stored on the blockchain [91], [126].

5) MANUAL ACTORS

On the opposite side of the autonomous actors, some data markets employ human actors to manually perform one of the identified roles. The most significant advantage of manual actors is that they can deal with the complexity that arises from the context-dependent nature of the problems discussed above [122]. The main downside of manual actors, however, is the lack of scaling when compared to automated processes [129]. This literature review has identified three types of problems that manual actors are addressing in the context of data markets. Firstly, manual data brokers search for highly specialised data products that may or may not be publicly offered on the data market registry [8], [35] or, alternatively, for requests for data products that they can provide [130], [131]. Secondly, the most common manual actors in the literature review are the data transformers, who create new data products from existing ones; either by adding new value or insights, standardising data products or by aggregating different data products into new ones [35], [84], [126]. Finally, manual data quality assessment relies on human actors to assess the quality of data products in scenarios where the data quality is hard to address in a standardised way [97], [101].

6) LOGGING

One of the most straightforward ways to improve trust in the ecosystem is to keep a log of all activities happening inside the ecosystem. In particular, the log increases the transparency of other solutions, processes and functionalities, which allows the different actors to perceive each other's behaviour and assess the trustworthiness of the different actors, data products, and the data market as a whole [116], [128].

7) IDENTITY MANAGEMENT MECHANISM

Providing an identity management mechanism enables the identity provider to keep track of the identities of the different actors in the data market ecosystem. Identity management makes it possible to hold actors personally responsible or enforce legal agreements, and this, in turn, discourages strategic behaviour that could be the result of anonymity [29], [31]. When implemented decentrally, a registry of identities can be tracked by an autonomous smart contract [132], [133].

8) NODE PARTICIPANT MANAGEMENT

In some decentral data markets, identity management enables more proactive management of participants. Node participant management is usually performed by a certificate provider, who certifies which actors can host the nodes that make up the decentralised network. Controlling the nodes that participate in the network is especially useful for consortium-based data markets that are trying to control who can participate and to what capacity [32], [83].

9) DATA & METADATA MODELS

Creating standards for (meta)data models is one of the most proliferate and versatile solutions discussed in the reviewed literature; in general, the goal of these models is to standardise either the metadata describing the data asset or the data asset itself. For example, a data broker could provide a standard metadata model to address the problem of data brokering because it enables a more straightforward and effective comparison between data products, allowing data consumers to find the most suitable data product [33], [65]. This practice is widespread in domain-specific data markets such as in data markets for cars [82], or data markets for IoT data [75]. In a similar vein, metadata models and data models that standardise data assets make it much easier for a quality assessor to compare different quality aspects of data products [128], thus addressing the problem of data quality. In addition to facilitating comparison between data products, (meta-)data models can help make data products more integrable; in fact, many (meta) data modelling standards have been designed specifically to link different data products in a logical way [87], [134]. These links, in turn, address some of the problems of data transformation, namely the challenges associated with data integrability. Moreover, if data assets are formatted in a standardised model, this makes it to develop standardised data transformation techniques [28].

10) AUTOMATED DATA PRODUCT TRANSFORMATION

Some data markets propose that a data broker implement an automated data transformation tool to generate standardised data product metadata automatically [28]. As was mentioned above, this can solve the challenge of data brokering because it makes it easier for data providers to compare data products and find the most suitable one. Alternatively, automated data product transformation techniques can be applied to create wholly new and improved data products, such as combining different data products to fit the data consumers need [34] or creating insights from raw data [35].

11) SPECIALISED QUERYING MECHANISM

Querying mechanisms are tools that allow data providers to search amongst available data products and find the most relevant data for them. In the domain of data markets, querying mechanisms are designed by the data broker so the data consumer can query data products and solve the problems of

data brokering. There are two levels where querying mechanisms can be applied: querying can happen on a data level or a metadata level. In the first case, the data consumer is trying to find relevant data *inside* one or more data products and data querying is often combined with a pay-per-query data pricing approach, e.g., [135]. Alternatively, querying can happen on a metadata level to query for the right data product(s) from amongst all the data products, e.g., [136]. These querying mechanisms are, of course, not mutually exclusive, and some data markets combine both types [137].

12) CROWD-SOURCING

Another way to address the problems of data brokering is by giving the data consumer the ability to employ crowd-sourcing to get the data they desire. With crowd-sourcing, the data consumer specifies some requirements for data, which is then created or gathered by data providers specifically to fit these requirements. One straightforward example of crowd-sourcing to solve the data brokering problem is the previously mentioned solution by Halpin and Lykourentzou [61] which allows data consumers to hire domain experts to jointly create specific data products on request. Another, more common, approach comes from the domain of IoT and is known as *sensing-as-a-service*. With sensing-as-a-service, the data consumer does not actually consume a data product but instead gains access to, or control over, one or more sensors, which in turn can be used to create the desired data [93], [128]. In both cases, the data broker should provide a vocabulary for the data consumer to express their requirements and for the data providers to express the data they can create.

13) RECOMMENDER SYSTEM

Even though only one work in the literature review considered a recommender system [138], we still consider it as a solution category because recommender systems are prevalent in other digital distribution platforms [139]. Therefore, it stands to reason that similar solutions would apply to solving data brokering problems as well. A recommender system designed by a data broker can support the decision-making process of a data consumer by learning about data consumer preferences and identifying patterns of similar consumers. For new data consumers, the recommender system could ask guiding questions and propose potentially useful data products [138].

14) DATA TRANSFORMATION ENVIRONMENT

Some data markets provide an environment on their platform where data transformers can manually transform data products without them ever leaving the data markets. This approach is better suited to centralised data markets than decentralised ones because it requires a centralised authority that controls the data products e.g., [50], [131]. However, some decentralised approaches for IoT exist, which put the burden of transformation on the data provider: An example is the use of *edge computing*, whereby the data is transformed through a platform that uses hardware provided by the data provider [52]. Another approach for a decentralised data transformation environment is proposed by Pillmann *et al.*, who provide a transformation environment to help data providers transform data assets as they create data products [82].

15) USAGE POLICY MANAGEMENT

A crucial step in data exchange is to agree on the responsibilities of both the data provider and the data consumer before the data product can be consumed. This agreement, or *usage policiy*, can be provided unilaterally by the data provider [7], [64] or data consumer [100], the data market provider can dictate it as part of the data market policy [60], or it can be negotiated to fit the specific needs and requirements of the data provider and data consumer [11], [20]. Agreements for data consumption come in many different forms, such as service level agreements (SLA's) [67], legal contracts [7] or they are expressed in specifically designed policy description languages [32]. Sometimes, they are simply recorded, in which case enforcement is the responsibility of individual data providers and consumers who can try to call on legal enforcement if all else fails [140]. In other cases, the data market might provide a method for enforcing policies internally or resolving disputes that arise [115], [141]. Usage policies help solve three types of problems: Firstly, they address the problems of transaction enforcement by formalising the conditions under which the consumption happens. Secondly, usage policies enable the data provider to state their requirements for data consumption, aiding in maintaining and enforcing data governance. Finally, usage policies can help address the problems of data quality by specifying which quality aspects should be provided by the data provider, particularly for the service-level quality facet.

16) ORIGIN TRACING MECHANISM

Several techniques exist to prove the origin of a data product; most commonly, some version of *digital watermarking* is applied, which involves hiding an (often encrypted) proof of ownership inside the data product [93], [126]. However, notable exceptions to the watermarking approach exist, such as the GeoHex method developed by Özyilmaz *et al.*, whereby IoT device data providers can vote on whether other IoT devices are close to them and hence, where they are operating from [99]. Another approach is discussed by Niya *et al.* and is based on challenge-response pairs to prove authenticity and ownership of data products. Data origin can be a (context-based) quality requirement in-and-of-itself, for example, when the geographic location of the data product is relevant [99]. Moreover, tracing data origins can be used to prove the preservation of *data integrity*, i.e., that data has not been altered or tampered with [95]. The advantage of using origin tracing mechanisms for maintaining data governance is that it allows data providers to prove where the data product came from and, consequently, which restrictions apply to it.

17) QUALITY METRICS

One of the most straightforward ways to address the challenges of determining data quality is to develop quality metrics. These quality metrics are properties of data products that can be measured in a standardised manner that reflect the different aspects of data quality (mainly from the service-level and content-based facets). Once these metrics have been defined, they can be used in several ways to ensure data consumers get access to high-quality data products. For example, (automated) services can be deployed that verify the quality of data products for data consumers [27], (automated) data transformation techniques can be used to improve the scores of data products according to the data quality metrics [33] or actors can manually check whether quality metrics adhere to the advertised quality [20].

18) CERTIFICATION FRAMEWORK

A less popular approach to ensuring data quality than developing quality metrics is using certification frameworks, which occurred only three times in the literature review. By relying on the certification of actors in the data market, the focus is shifted from quality properties of data products towards quality properties of data providers (and sometimes other actors as well). The implicit idea behind this shift in focus is that the certificates act as a proxy for data quality because certified data providers will provide high-quality data. Because of this, certification can be optional and simply be used as an indication of quality [96], or they can be a core feature of the data market, which forces data providers to conform to the standards necessary for certification or risk losing the right to participate [75], [142].

19) REPUTATION

Using reputation as a proxy for data quality is a common technique used in markets for digital goods [143], [144] and it is therefore not surprising that this is a standard method for assessing the quality of data products. The reputation of different data products and data providers can be ascertained from their performance on the data market, in which case it is usually combined with extensive logging as described above [145], [146]. More proactive approaches exist as well, however, such as the solicitation of direct feedback from data consumers [91], [99] and the use of *token-curated-registries*, whereby data providers maintain and curate registries of data products and new data providers have to convince others of the quality of their data product in order to be admitted to the registry [97], [141].

20) COMPUTE-TO-DATA

Compute-to-data has already been discussed as an access method in section III-A, but it is also a solution that addresses problems of transaction enforcement, as well as maintaining data governance. With compute-to-data, rather than handing over the data to the data consumer for consumption in their environment, the consumer sends code to manipulate the data and receives results of computation that generally do not expose the underlying data. In its simplest form, the data provider sends the code directly to the data provider, who either provides an environment for execution of the code or runs the code manually themselves [19], [52]. However, since code can also be considered a data product, the data consumer might not be willing to share their code for the same reasons the data provider is not. In such a case, a trusted third party can be charged with providing an isolated environment where the code can be executed on the data [31], [34]. A special case of compute-to-data is the use of *multi-party computation* whereby code is run on multiple data products in multiple physical locations, and an aggregate result is obtained without revealing information from the individual data products [147].

21) ACCESS CONTROL MECHANISM

Access control mechanisms are the most prolific and simultaneously one of the most versatile solutions identified in the literature review. When data markets are designed, the access control mechanism allows the data provider to control and specify the manner in which the data consumer can access their data products [100]. Some access control mechanisms, such as compute-to-data and the use of a centralised clearinghouse, have already been discussed in this section. However, different manifestations exist that use other solutions to allow the data provider the required control for accessing their data products; these are discussed below. Sometimes, the data is encrypted, and access is granted relatively straightforwardly because the data cannot be used without the transfer of an encryption key, which can be transferred securely in a peer-to-peer manner [127]. This method is easy to implement and allows for the use of cryptographic proof-of-knowledge algorithms, whereby the data consumer can verify specific properties of the encrypted data without needing to decrypt it [90]. An obvious downside of making data readily available is that, even if it is sufficiently encrypted, technological progress might make decryption of data products feasible at a later stage [16]. Another straightforward approach combines access control management with usage control policies by including access policies and leveraging the methods already in place for enforcing usage policies also to enforce access policies [32], [111]. This approach offers fine-grained control over access policies but requires more investments into the infrastructure necessary for defining, creating and managing the required policies. Finally, about 16% of all identified works make use of autonomous actors that act as an access control mechanism. These implementations vary from custom-made smart contracts (e.g., [29], [90] to the use of existing platforms such as the previously mentioned IOTA [69], [72], [97].

22) ANONYMISATION TECHNIQUES

Several works consider anonymisation techniques, which aim to transform data products so that they can still be usefully consumed without giving away information that the data

provider wants to keep confidential (e.g., personal information protected by privacy legislation). Three approaches can be distilled from the works in this literature review. The first approach is *k-anonimity*, which works by aggregating data and eliminating personally-identifying features until each individual whose information is in the data cannot be distinguished from that of at least $k - 1$ other individuals [148]. This technique is often used in data markets for personal data where a trusted intermediary aggregates data and ensures k-anonymity [71], [115]. Another technique that similarly requires aggregated data in order to be applied is *differential privacy*. Differential privacy introduces just enough noise into query results to the point where the results no longer depend on the information of any single entry in a database [149], [150]. Finally, some data markets rely on *obfuscation*, a technique whereby information (such as code that is being used in a compute-to-data solution) is transformed, so it is difficult for humans to read [31], [100].

23) CENTRAL CLEARINGHOUSE

Central clearinghouses work in much the same manner as autonomous clearinghouses, with the exception that they do not rely on blockchain- and smart contract technology. Instead, a platform provider in a centralised data market can implement the necessary software for holding on to data and potential payment and enforce the transaction exchange as necessary [87], [151], [152]. By using a central clearinghouse over an autonomous one, a data market can avoid some of the challenges that come with smart contracts, such as flexibility, maintainability and scalability that come with avoiding smart contracts [153]. Nevertheless, central clearinghouses rely heavily on the trust of the data providers (and, to a lesser extend, data consumers) to correctly handle data products and potential payments.

24) MIDDLEWARE

When data providers and data consumers are not presumed to make the exchange directly themselves, and there is also no clearinghouse (central or autonomous) to act as a middleman, the platform provider can instead choose to provide middleware as a means of connecting the data provider and data consumer. Middleware is an oft-used solution for providing communication between separate nodes in decentralised architectures, and data markets are no exception [11], [80], [116].

V. CONCLUSION

The results presented above give an overview of the state-of-the-art in academic research over a broad scope of design options, application domains, problems and solutions and their relations. Based on these results, we present a group of five types of data markets, each of which frequently occurs in literature. For each data market type, we identify a coherent set of *best practices*, which are the most commonly proposed roles to emphasise, problems to focus on, and solutions that address these problems. These best practices guide design considerations and serve as a starting point for software architects and engineers looking to design a data market. Table 4 shows an overview of the types of data markets and the identified best practices, each of which is discussed in further detail below. It is important to note that not all of these types are mutually exclusive: it is possible to find a data market that is both a specialist and an aggregator. In those cases, best practices from both types can be combined to help with the design process of the data market.

A. THE GENERALIST

The generalist data market is characterised by its facilitation of heterogeneous data across multiple domains. Generalist data markets can also manifest inside a single large company or organisation with many departments, for example, when a company moves from a data warehouse solution to a data mesh solution [94]. The heterogeneity of the data products is due, in part, to the fact that there are many different data providers, each with their own data & metadata models that provide data for many different consumers with different use-cases.

At the same time, the heterogeneous nature of the data products makes it hard to apply general solutions to problems such as maintaining data governance, transforming data or assessing data quality. Instead, general data markets should focus on how to facilitate data brokering and transaction enforcement in a way that is independent of data product structure.

Central clearinghouses work well in this context because they provide an environment for data providers to upload their products and naturally set minimum requirements for what can be uploaded and thus exchanged. For addressing data brokering, most general data markets choose a general metadata standard along with a specialised querying mechanism that can distinguish between similar but different data products. Finally, since the heterogeneity of the data products makes it impossible to leverage specialised solutions, manual actors become valuable solutions to problems other than data brokering and transaction enforcement. For example, manual data transformers can buy similar different data products and then turn around and provide the aggregated information as a new data product.

B. THE SPECIALIST

The specialist data market is, in many aspects, the opposite of the generalist data market as they focus on one (or a few) types of strongly homogenised data products that originate from a single domain. Moreover, specialist data markets also allow for many-to-one matching (e.g., for crowdsourcing data markets [61]).

Since specialist data products always come from a single domain, the critical problems and corresponding central roles can vary significantly: personal data leads to an emphasis on data governance, spatial data leads to an emphasis on data brokering, machine learning data leads to an emphasis on data transformation. Nevertheless, homogeneous data

TABLE 4. Five types of data markets frequently occurring in literature and best practices designing them. Each type can be identified by its defining characteristics and typically focuses on either one, or multiple domains. Although all roles should be considered in data market design, different types emphasise one or more roles over others. These actors address the problems that are crucial to each type's success, and literature suggests some typical solutions that can be considered best practices for addressing these problems in the corresponding context. Finally, for each type of data market, we also provide some example works that can serve as a starting point for further reading.

	Generalist	Specialist	Industry Exchange	Enabler	Aggregator
Defining Characteristics	- Heterogeneous data - Domain agnostic - Many-to-many matching	- Homogeneous data - Single Domain	- Providers & Consumers are companies/organisations - Data from one domain but heterogeneous structure - Decentral architecture - Consortium-owned - Specialised Software	- Many-to-many matching - Small data products	Many-to-one + one-to-many matching - Extensive control of all processes - Monopoly
Central Roles	- Data Broker - Clearinghouse	- Domain Dependent - Data Transformer	- Infrastructure Provider - Identity Provider - Certificate Provider	- Clearinghouse - Infrastructure Provider	- Data Transformer
Critical Problems	- Data Brokering - Transaction Enforcement	Domain Dependent - Data Transformation	- Data Governance	- Data Transformation - Transaction Enforcement	- Data Transformation - Data Governance
Typical Solutions	- Central Clearinghouse - Specialised Querying Mechanism - Manual Actors	- Quality Metrics - Automation - Compute-to-data	- Identity Management Node Participation Management - Certification Framework - Usage Policies	- Middleware - Central/automated clearinghouse - Manual transformation - Transformation Environment	- Anonymisation Techniques Data & - Metadata Models - Usage Policies
Example Works	Hayashi & Ohsawa [131], Spiekermann [15], Nguyen & Won [154]	Ahmed & Shabani [8], Sakr [65], Sajan et al. [52], Alsharif & Nabil [90]	Llewelyn et al. [50], Munoz-Arcentales et al. [111], Pillman et al. [82], Radhakrishnan & Das [96]	Cao et al. [11], Jeong et al. [87], Figueredo et al. [88], Perera et al. [20]	Eng et al. [60], Niu et al. [36], Thomas & Leiponen [12], Liang et al. [155]

products are usually created for a specific, well-known use case. Because of this, specialist data markets can make assumptions on how data will be used by consumers and should focus on alleviating problems associated with data transformation.

The homogeneous nature of the data products implies the existence of standard data & metadata models for the data products in the domain, and so, the data market should not have to spend effort creating these. Instead, these standards enable automated operations, such as automated transformation or a tailored data transformation platform. Similarly, quality assurance can easily be addressed by making use of domain-driven quality metrics. Finally, compute-to-data techniques rely heavily on knowing the structure of the underlying data products [52] which is known in specialist data markets, and these can therefore be used.

C. INDUSTRY DATA EXCHANGE

In an industry data exchange, data providers and data consumers are (large) companies, usually from the same domain, that want to exchange data products, for example, to optimise a manufacturing supply chain [9]. Generally, these companies provide and consume significantly more data products than individuals would. Consequently, they are also willing to invest in the infrastructure necessary to provide the platform and develop specialised software to facilitate the exchange. This means that industry data exchanges generally follow a decentral architecture and are owned by a consortium of data providers and consumers that work with a many-to-many matching model.

There are two reasons why data governance is the most emphasised problem for industry data exchanges. First, most companies are starting to view their data as valuable assets and want to control how, when and for what purpose ownership of the data can be transferred [1], [2]. Secondly, it has long been standard practice for companies to collect data from their customers and users [156] which is guided by privacy policies that need to be also respected when creating and exchanging data products.

The best practices identified in literature for addressing these problems are solid identity management, node participation management and certificate management. These solutions can then be combined with extensive access- and usage policies that can be enforced on the specialised software that the consortium infrastructure providers run.

D. THE ENABLER

In contrast to industry exchanges, *enabler* data markets look to enable a large number of individual data providers with relatively little data per provider that do not have access to the same resources that companies do. Enabler data markets are widespread in the IoT domain, where many different actors own one or a few sensors that produce relatively little data (e.g., smart health, smart home, smart city, smart mobility).

Although any of the problems identified in section IV-C can be relevant for enabler data markets, two stand out in particular. Firstly, all enabler data markets share an emphasis on addressing transaction enforcement on behalf of the data providers and data consumers. This emphasis on enabler data markets always taking on the role of clearinghouse can be explained by the data providers and data consumers limited resources preventing them from taking on the role themselves. Secondly, because enabler data markets rely on many small data providers with relatively little data per data product, it is presumed that the data consumer wants to aggregate data from multiple data providers. Thus data transformation is always one of the main problems for enabler data markets.

In order to address the problems of transaction enforcement and data transformation in the context of enabler data markets, several solutions appear as best practices in the works in this literature review. First of all, enabler data markets provide explicit tools for transaction enforcement, whether middleware or a (central or autonomous) clearinghouse. When it comes to addressing data transformation, the large number of individual data providers makes it hard to guarantee metadata or schema standards on the data products. Therefore, the best practice for addressing data transformation in the context of enabler data markets is to rely on manual data transformers. If possible, this approach can be enhanced by providing a data transformation environment.

E. THE AGGREGATOR

The final data market type for which best practices are identified is the aggregator data market. In an aggregator data market, the platform provider acts as a private buyer consuming data from many data providers through a many-to-one matching system, transforms and aggregates the data, and then proceeds to act as a private seller that provides the transformed data to many data consumers in a one-to-many matching system. This type of data market is characterised by the high level of control that the platform provider has over the data products and the platform itself. On the many-to-one side, the platform provider creates the platform on which the data is created (usually in exchange for a service such as with social media platforms [157], smart sensors [60], or app usage [156]. On the many-to-one side, the platform provider will create standard access points (usually an API) for consuming the aggregated data.

Based on the work considered in this literature review, two main problems have been identified that can be focused on as a best practice: data transformation and data governance. Data transformation is critical because the goal of the platform provider is to sell valuable data, which it gains through aggregating and transforming data from its providers. On the other hand, maintaining data governance is important in so far as that the data that is collected is usually private data prior to aggregation.

There are several best practices for aggregator data markets. Usually, the data collection is governed by policies

such as End User License Agreements (EULAs) that the data providers have to agree with to use the services provided by the platform provider. Moreover, the high level of control that the platform provider has over the data allows for a high level of standardisation and automation. Finally, aggregation lends itself well to being combined with anonymisation techniques to facilitate data privacy concerns for data providers.

VI. DISCUSSION

In contrast to previous surveys the present study explicitly considered data markets as IT artefacts with technical implications and manifestations across different domains and contexts. As a result three main contributions were made. First, we presented an overview on the academic literature on data markets, including a comparison with existing market trends. Second, we identified important application domains and contexts where data markets are being put into practice. This allowed us to study various architectural designs and manifestations of data markets and list the key concepts that persist across different domains. Finally, we provided taxonomies of both design problems for data markets and the solutions found in the literature to address these problems.

The conclusion of this paper describes five common types of data markets described in the literature, including a set of best practices applied each type of data market. These types, best practices and example works from the literature may serve as a starting point for software architects and engineers when designing data markets. Furthermore, our review addresses the lack of common definitions, design standards, and terminologies in data market development that has been pointed out previously [1], [24]–[26].

This survey may be affected by the following threats to validity. As in many literature reviews the classification of studies is prone to subjectivity. However, the dimensions and values used in this review are based on both previous literature (reviews) on data markets and well-established methodology, namely grounded theory. Similarly, the paper selection might be considered subjective. To address this, the selection is the result of a formal protocol and relies on two reviewers who resolved classification disagreements through extensive discussion. Additionally, an inter-rater reliability criterion for the inclusion assessment was calculated and indicated a weak bias by the reviewers. Finally, while the search string used in our survey is relatively simple and can easily be linked to a large set of articles, it yield good results. The included papers allowed us to answer the proposed research questions, derive taxonomies of problems and solutions in data market design, and identify a set of five common types of data markets described in the literature.

The results of our survey open up future work that can use the conclusive overview on typical data market types as a base for the development of a pattern language, consisting of architectural patterns for data market design. It is recommended to study specific use cases of these design patterns together with the industry to further evince the practical value of such patterns and identify remaining relevant research challenges. In addition, the extensive coverage of different dimensions across domain boundaries opens up the way for cross-pollination of ideas for different types of data markets. The problem- and solution taxonomies, together with the common types of data markets discussed in the conclusion can serve as a sort of checklist for future works that are intent on presenting complete data market architectures.

ACKNOWLEDGMENT

The authors would like to thank Dr. Gemma Catalino for the fruitful discussions and useful suggestions that helped shape this paper and Dr. Indika Kumar for his contributions in formatting the paper.

REFERENCES

[1] F. Stahl, F. Schomm, G. Vossen, and L. Vomfell, "A classification framework for data marketplaces," *Vietnam J. Comput. Sci.*, vol. 3, no. 3, pp. 137–143, Aug. 2016.

[2] H. Zech, "Data as a tradeable commodity–implications for contract law," in *Proc. 18th EIPIN Congr., New Data Economy Between Data Ownership, Privacy Safeguarding Competition*, Forthcoming 2017.

[3] D. Reinsel, J. Gantz, and J. Rydning, "Data age 2025: The digitization of the world from edge to core," Seagate, Cupertino, CA, USA, Tech. Rep. Doc US44413318, Nov. 2018, p. 28. [Online]. Available: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[4] V. Koutsos, D. Papadopoulos, D. Chatzopoulos, S. Tarkoma, and P. Hui, "Agora: A privacy-aware data marketplace," in *Proc. IEEE 40th Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Nov. 2020, pp. 1211–1212.

[5] L. Rodríguez-Mazahua, C.-A. Rodríguez-Enríquez, J. L. Sánchez-Cervantes, J. Cervantes, J. L. García-Alcaraz, and G. Alor-Hernández, "A general perspective of big data: Applications, tools, challenges and trends," *J. Supercomput.*, vol. 72, no. 8, pp. 3073–3113, 2016.

[6] S. Sharma, "Rise of big data and related issues," in *Proc. 12th IEEE Int. Conf. Electron., Energy, Environ., Commun., Comput., Control (INDICON)*, Dec. 2016, pp. 1–6.

[7] P. Tzianos, G. Pipelidis, and N. Tsiamitros, "Hermes: An open and transparent marketplace for IoT sensor data over distributed ledgers," in *Proc. IEEE Int. Conf. Blockchain Cryptocurrency (ICBC)*, May 2019, pp. 167–170.

[8] E. Ahmed and M. Shabani, "DNA data marketplace: An analysis of the ethical concerns regarding the participation of the individuals," *Frontiers Genet.*, vol. 10, pp. 1–6, Nov. 2019.

[9] G. Shaabany, M. Grimm, and R. Anderl, "Secure information model for data marketplaces enabling global distributed manufacturing," *Proc. CIRP*, vol. 50, pp. 360–365, Jan. 2016, doi: 10.1016/j.procir.2016.05.003.

[10] S. Schlarb, R. Karl, R. King, T. J. Lampoltshammer, L. Thurnay, B.-P. Ivanschitz, and V. Mireles, "Using blockchain technology to manage membership and legal contracts in a distributed data market," in *Proc. 6th Int. Conf. Softw. Defined Syst. (SDS)*, Jun. 2019, pp. 272–277.

[11] T.-D. Cao, T.-V. Pham, Q.-H. Vu, H.-L. Truong, D.-H. Le, and S. Dustdar, "MARSA: A marketplace for realtime human sensing data," *ACM Trans. Internet Technol.*, vol. 16, no. 3, pp. 1–21, Aug. 2016.

[12] L. D. W. Thomas and A. Leiponen, "Big data commercialization," *IEEE Eng. Manag. Rev.*, vol. 44, no. 2, pp. 74–90, Jul. 2016.

[13] P. Koutroumpis, A. Leiponen, and L. D. W. Thomas, "Markets for data," *Ind. Corporate Change*, vol. 29, no. 3, pp. 645–660, 2020.

[14] F. Stahl, F. Schomm, L. Vomfell, and G. Vossen, "Marketplaces for digital data: Quo vadis?" Eur. Res. Center Inf. Syst., Münster, Germany, ERCIS Work. Paper 24, 2015.

[15] M. Spiekermann, "Data marketplaces: Trends and monetisation of data goods," *Intereconomics*, vol. 54, no. 4, pp. 208–216, Jul. 2019.

[16] G. Ishmaev, "The ethical limits of blockchain-enabled markets for private IoT data," *Philosophy Technol.*, vol. 33, no. 3, pp. 411–432, Sep. 2020.

[17] S. Barns, "Smart cities and urban data platforms: Designing interfaces for smart governance," *City, Culture Soc.*, vol. 12, pp. 5–12, Nov. 2018, doi: 10.1016/j.ccs.2017.09.006.
[18] A. Brandão, H. S. Mamede, and R. Gonçalves, "Trusted data's marketplace," in *New Knowledge in Information Systems and Technologies*, vol. 930. Cham, Switzerland: Springer, 2019, pp. 460–472.
[19] C. Perera, C. Liu, R. Ranjan, L. Wang, and A. Y. Zomaya, "Privacy-knowledge modeling for the Internet of Things: A look back," *Computer*, vol. 49, no. 12, pp. 60–68, Dec. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7756262/
[20] C. Perera, S. Y. L. Wakenshaw, T. Baarslag, H. Haddadi, A. K. Bandara, R. Mortier, A. Crabtree, I. C. L. Ng, D. McAuley, and J. Crowcroft, "Valorising the IoT databox: Creating value for everyone," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 1, p. e3125, Jan. 2017.
[21] A. Muschalle, F. Stahl, A. Löser, and G. Vossen, "Pricing approaches for data markets," in *Proc. Int. Workshop Bus. Intell. Real-Time Enterprise*, in Lecture Notes in Business Information Processing, vol. 154, 2013, pp. 129–144.
[22] N. Golrezaei and H. Nazerzadeh, "Pricing schemes for metropolitan traffic data markets," in *Proc. 3rd Int. Conf. Data Manage. Technol. Appl. (DATA)*, 2014, pp. 266–271.
[23] S. A. Fricker and Y. V. Maksimov, "Pricing of data products in data marketplaces," in *Proc. Int. Conf. Softw. Bus.*, in Lecture Notes in Business Information Processing, vol. 304, 2017, pp. 49–66.
[24] F. Schomm, F. Stahl, and G. Vossen, "Marketplaces for data: An initial survey," *ACM SIGMOD Rec.*, vol. 42, no. 1, pp. 15–26, May 2013.
[25] F. Stahl, F. Schomm, and G. Vossen, *Data Marketplaces: An Emerging Species* (Frontiers in Artificial Intelligence and Applications). Amsterdam, The Netherlands: IOS Press, 2014, pp. 145–158.
[26] B. Varghese, M. Villari, O. Rana, P. James, T. Shal, M. Fazio, and R. Ranjan, "Realizing edge marketplaces: Challenges and opportunities," Dec. 2018, *arXiv:1812.01344*.
[27] P. Sharma, S. Lawrenz, and A. Rausch, "Towards trustworthy and independent data marketplaces," in *Proc. 2nd Int. Conf. Blockchain Technol.*, Mar. 2020, pp. 39–45.
[28] Y. Kim and E.-N. Huh, "Study on user customized data service model for improving data service reliability," in *Proc. 11th Int. Conf. Ubiquitous Inf. Manage. Commun. (IMCOM)*, Jan. 2017, pp. 1–8.
[29] M. J. M. Chowdhury, M. S. Ferdous, K. Biswas, N. Chowdhury, A. Kayes, P. Watters, and A. Ng, "Trust modeling for blockchain-based wearable data market," in *Proc. IEEE Int. Conf. Cloud Comput. Technol. Sci. (CloudCom)*, Dec. 2019, pp. 411–417.
[30] Y. Zeng and Y. Ohsawa, "Re-discover values of data using data jackets by combining cluster with text analysis," *Proc. Comput. Sci.*, vol. 112, pp. 2195–2203, Jan. 2017, doi: 10.1016/j.procs.2017.08.111.
[31] Z. Wang, Z. Zheng, W. Jiang, and S. Tang, "Blockchain-enabled data sharing in supply chains: Model, operationalization, and tutorial," *Prod. Oper. Manage.*, vol. 30, no. 7, pp. 1965–1985, Jul. 2021.
[32] B. Otto, S. Steinbuß, A. Teuscher, and S. Lohmann, "IDSA reference architecture model," Int. Data Spaces Assoc., Dortmund, Germany, Tech. Rep., 2019. [Online]. Available: https://internationaldataspaces.org/download/16630/
[33] K. Nagorny, S. Scholze, M. Ruhl, and A. W. Colombo, "Semantical support for a CPS data marketplace to prepare big data analytics in smart manufacturing environments," in *Proc. IEEE Ind. Cyber-Physical Syst. (ICPS)*, May 2018, pp. 206–211.
[34] R. C. Fernandez, P. Subramaniam, and M. J. Franklin, "Data market platforms: Trading data assets to solve data problems," *Proc. VLDB Endowment*, vol. 13, no. 12, pp. 2150–8097, 2020.
[35] J. Attard, F. Orlandi, and S. Auer, "Exploiting the value of data through data value networks," in *Proc. 10th Int. Conf. Theory Pract. Electron. Governance*, Mar. 2017, pp. 475–484.
[36] C. Niu, Z. Zheng, F. Wu, X. Gao, and G. Chen, "Achieving data truthfulness and privacy preservation in data markets," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 1, pp. 105–119, Jan. 2019.
[37] H. Bastian, P. Glasziou, and I. Chalmers, "Seventy-five trials and eleven systematic reviews a day: How will we ever keep up?" *PLoS Med.*, vol. 7, no. 9, Sep. 2010, Art. no. e1000326.
[38] E. Coren and M. Fisher, "The conduct of systematic research reviews for SCIE knowledge reviews," Social Care Inst. Excellence, London, U.K., Tech. Rep., 2006. [Online]. Available: http://lx.iriss.org.U.K./sites/default/files/resources/The conduct.pdf
[39] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering, version 2.3," EBSE, Softw. Eng. Group, School Comput. Sci. Math., Keele Univ., Keele, U.K., Dept. Comput. Sci., Univ. Durham, Durham, U.K., Tech. Rep. EBSE-2007-01, 2007.
[40] D. Tranfield, D. Denyer, and P. Smart, "Towards a methodology for developing evidence-informed management knowledge by means of systematic review," *Brit. J. Manage.*, vol. 14, no. 3, pp. 207–222, Sep. 2003.
[41] O. Ali, M. Ally, Clutterbuck, and Y. Dwivedi, "The state of play of blockchain technology in the financial services sector: A systematic literature review," *Int. J. Inf. Manage.*, vol. 54, Oct. 2020, Art. no. 102199. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0268401219310928
[42] B. Kitchenham, "Procedures for performing systematic reviews," NICTA, Keele Univ., Keele, U.K., Tech. Rep. 0400011T.1, 2004.
[43] R. J. Wieringa, *Design Science Methodology: For Information Systems and Software Engineering*. Berlin, Germany: Springer, 2014.
[44] A. Hevner and S. Chatterjee, *Design Research in Information Systems: Theory and Practice*, vol. 28, S. Ramesh and S. Voß, Eds. New York, NY, USA: Springer, 2010.
[45] P. Y. Martin and B. A. Turner, "Grounded theory and organizational research," *J. Appl. Behav. Sci.*, vol. 22, no. 2, pp. 141–157, 1986.
[46] A. Straus and J. Corbin, "Grounded theory methodology: An overview," in *Handbook of Qualitative Research*, 1st ed. New Delhi India: Sage, 1994, pp. 273–285.
[47] A. E. Roth, "The economist as engineer: Game theory, experimentation, and computation as tools for design economics," *Econometrica*, vol. 70, no. 4, pp. 1341–1378, Jul. 2002.
[48] A. E. Roth, "What have we learned from market design?" *Econ. J.*, vol. 118, no. 527, pp. 285–310, Mar. 2008.
[49] B. de Best and P. Huijbers, *Cloud SLA*, 1st ed., Zutphen, The Netherlands: Leonon Media, (in Dutch), 2014.
[50] T. Llewellynn, M. M. Fernández-Carrobles, O. Deniz, S. Fricker, A. Storkey, N. Pazos, G. Velikic, K. Leufgen, R. Dahyot, S. Koller, G. Goumas, P. Leitner, G. Dasika, L. Wang, and K. Tutschku, "BONSEYES: Platform for open development of systems of artificial intelligence: Invited paper," in *Proc. Comput. Frontiers Conf.*, May 2017, pp. 299–304.
[51] W. Wingerath, N. Ritter, and F. Gessert, *Real-Time & Stream Data Management: Push-Based Data in Research & Practice*. Cham, Switzerland: Springer, 2019.
[52] K. K. Sajan, G. S. Ramachandran, and B. Krishnamachari, "Enhancing support for machine learning and edge computing on an IoT data marketplace," in *Proc. AIChallengeIoT*, 2019, pp. 19–24.
[53] S. Golder, Y. K. Loke, and L. Zorzela, "Comparison of search strategies in systematic reviews of adverse effects to other systematic reviews," *Health Inf. Libraries J.*, vol. 31, no. 2, pp. 92–105, Jun. 2014.
[54] B.-J. Butijn, D. A. Tamburri, and W.-J.-V. D. Heuvel, "Blockchains: A systematic multivocal literature review," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–37, May 2021.
[55] R. Alexandre-Benevant, G. A. Gregorio, J. G. De Dios, and A. Alonso-Arroyo, "Sources of bibliographic information. Rationale for conducting a literature search," *Acta Pediatrica Espanola*, vol. 69, no. 3, pp. 131–136, 2011.
[56] I. Tahamtan, A. S. Afshar, and K. Ahamdzadeh, "Factors affecting number of citations: A comprehensive review of the literature," *Scientometrics*, vol. 107, no. 3, pp. 1195–1225, 2016.
[57] M. Wang, G. Yu, and D. Yu, "Effect of the age of papers on the preferential attachment in citation networks," *Phys. A, Stat. Mech. Appl.*, vol. 388, no. 19, pp. 4273–4276, 2009, doi: 10.1016/j.physa.2009.05.008.
[58] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Commun. Methods Measures*, vol. 1, no. 1, pp. 77–89, Apr. 2007.
[59] A. Straus and J. Corbin, *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. New Delhi India: Sage, 2014.
[60] K. Eng, D. Serrano, E. Stroulia, and J. Jaremko, "(Semi)automatic construction of access-controlled web data services," in *Proc. 28th Annu. Int. Conf. Comput. Sci. Softw. Eng.*, Oct. 2018, pp. 72–80. [Online]. Available: http://dl.acm.org/citation.cfm?id=3291291.3291300
[61] H. Halpin and I. Lykourentzou, "Crowdsourcing high-quality structured data," in *Proc. Annu. Int. Symp. Inf. Manage. Big Data*. Cham, Switzerland: Springer, 2018, pp. 304–319, doi: 10.1007/978-3-030-11680-4_29.
[62] L. Moor, L. Bitter, M. D. Prado, N. Pazos, and N. Ouerhani, "IoT meets distributed AI–deployment scenarios of bonseyes AI applications on FIWARE," in *Proc. IEEE 38th Int. Perform. Comput. Commun. Conf. (IPCCC)*, Oct. 2019, pp. 4–5.

[63] R. Xu, G. S. Ramachandran, Y. Chen, and B. Krishnamachari, "BlendSM-DDM: BLockchain-ENabled secure microservices for decentralized data marketplaces," in *Proc. IEEE Int. Smart Cities Conf. (ISC)*, Oct. 2019, pp. 14–17.

[64] A. Gerl and B. Meier, "Privacy in the future of integrated health care services—Are privacy languages the key?" in *Proc. Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, Oct. 2019, pp. 312–317.

[65] M. Sakr, "A data model and algorithms for a spatial data marketplace," *Int. J. Geograph. Inf. Sci.*, vol. 32, no. 11, pp. 2140–2168, Nov. 2018, doi: 10.1080/13658816.2018.1484124.

[66] S. Lawrenz, P. Sharma, and A. Rausch, "The significant role of metadata for data marketplaces," in *Proc. Int. Conf. Dublin Core Metadata Appl.*, 2019, pp. 95–101.

[67] M. Spiekermann, D. Tebernum, S. Wenzel, and B. Otto, "A metadata model for data goods," *Multikonferenz Wirtschaftsinformatik*, vol. 2018, pp. 326–337, Mar. 2018.

[68] (2018). *The Fi-Ware Project*. [Online]. Available: https://www.fiware.org

[69] S. Musso, G. Perboli, M. Rosano, and A. Manfredi, "A decentralized marketplace for M2M economy for smart cities," in *Proc. IEEE 28th Int. Conf. Enabling Technol., Infrastruct. Collaborative Enterprises (WETICE)*, Jun. 2019, pp. 27–30.

[70] Z.-J. Wang, C.-H.-V. Lin, Y.-H. Yuan, and C.-C.-J. Huang, "Decentralized data marketplace to enable trusted machine economy," in *Proc. IEEE Eurasia Conf. IoT, Commun. Eng. (ECICE)*, Oct. 2019, pp. 246–250.

[71] M. Zichichi, M. Contu, S. Ferretti, and V. Rodríguez-Doncel, "Ensuring personal data anonymity in data marketplaces through sensing-as-a-service and distributed ledger technologies," in *Proc. CEUR Workshop*, vol. 2580, 2020, pp. 1–16.

[72] S. Popov. (2018). *IOTA Whitepaper V1.4.3*. [Online]. Available: http://www.descryptions.com/Iota.pdf

[73] D. Ramel. (2016). *Microsoft Closing Azure DataMarket*. [Online]. Available: https://adtmag.com/articles/2016/11/18/azure-datamarket-shutdown.aspx

[74] A. S. Gilles. (2021). *What is Internet of Things (IoT)?* [Online]. Available: https://internetofthingsagenda.techtarget.com/definition/Internet-of-Things-IoT

[75] L. Sánchez, J. Lanza, J. Santana, R. Agarwal, P. Raverdy, T. Elsaleh, Y. Fathy, S. Jeong, A. Dadoukis, T. Korakis, S. Keranidis, P. O'Brien, J. Horgan, A. Sacchetti, G. Mastandrea, A. Fragkiadakis, P. Charalampidis, N. Seydoux, C. Ecrepont, and M. Zhao, "Federation of Internet of Things testbeds for the realization of a semantically-enabled multi-domain data marketplace," *Sensors*, vol. 18, no. 10, p. 3375, Oct. 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/10/3375

[76] D. Roman and G. Stefano, "Towards a reference architecture for trusted data marketplaces: The credit scoring perspective," in *Proc. 2nd Int. Conf. Open Big Data (OBD)*, Aug. 2016, pp. 95–101.

[77] S. Spiekermann and A. Novotny, "A vision for global privacy bridges: Technical and legal measures for international data markets," *Comput. Law Secur. Rev.*, vol. 31, no. 2, pp. 181–200, Apr. 2015, doi: 10.1016/j.clsr.2015.01.009.

[78] EU. (2016). *General Data Protection Regulation*. [Online]. Available: https://gdpr-info.eu/

[79] R. Bonta. (2018). *California Consumer Privacy Act*. [Online]. Available: https://www.oag.ca.gov/privacy/ccpa

[80] D. Grewe, M. Wagner, and H. Frey, "ICN-based open, distributed data market place for connected vehicles: Challenges and research directions," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, May 2017, pp. 265–270.

[81] J. Pillmann, B. Sliwa, J. Schmutzler, C. Ide, and C. Wietfeld, "Car-to-cloud communication traffic analysis based on the common vehicle information model," in *Proc. IEEE 85th Veh. Technol. Conf. (VTC Spring)*, Jun. 2017, pp. 1–5.

[82] J. Pillmann, C. Wietfeld, A. Zarcula, T. Raugust, and D. C. Alonso, "Novel common vehicle information model (CVIM) for future automotive vehicle big data marketplaces," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2017, pp. 1910–1915.

[83] B. G. Jeong, T. Y. Youn, N. S. Jho, and S. U. Shin, "Blockchain-based data sharing and trading model for the connected car," *Sensors*, vol. 20, no. 11, pp. 1–20, 2020.

[84] K. Figueredo, D. Seed, and V. Subotic, "Preparing for highly scalable and replicable IoT systems," *IEEE Internet Things Mag.*, vol. 3, no. 3, pp. 94–98, Sep. 2020.

[85] I. Docherty, G. Marsden, and J. Anable, "The governance of smart mobility," *Transp. Res. A, Policy Pract.*, vol. 115, pp. 114–125, Oct. 2018, doi: 10.1016/j.tra.2017.09.012.

[86] V. Albino, U. Berardi, and R. M. Dangelico, "Smart cities: Definitions, dimensions, performance, and initiatives," *J. Urban Technol.*, vol. 22, no. 1, pp. 3–21, 2015, doi: 10.1080/10630732.2014.942092.

[87] S. Jeong, S. Kim, and J. Kim, "City data hub: Implementation of standard-based smart city data platform for interoperability," *Sensors*, vol. 20, no. 23, pp. 1–20, 2020.

[88] K. Figueredo, D. Seed, and C. Wang, "A scalable, standards-based approach for IoT data sharing and eco-system monetization," *IEEE Internet Things J.*, early access, Sep. 9, 2020, doi: 10.1109/JIOT.2020.3023035.

[89] E. Hamilton. (2018). *What is Edge Computing: The Network Edge Explained*. [Online]. Available: https://www.cloudwards.net/what-is-edge-computing/

[90] A. Alsharif and M. Nabil, "A blockchain-based medical data marketplace with trustless fair exchange and access control," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, p. 5.

[91] J. S. Park, T. Y. Youn, H. B. Kim, K. H. Rhee, and S. U. Shin, "Smart contract-based review system for an IoT data marketplace," *Sensors*, vol. 18, no. 10, pp. 1–16, 2018.

[92] (2021). *Industry 4.0*. [Online]. Available: https://www.plattform-i40.de/PI40/Navigation/EN/Home/home.html

[93] J.-T. S. Altmann and Jörn, "Sensing as a service revisited: A property rights enforcement and pricing model for IIoT data marketplaces," in *Proc. Int. Conf. Econ. Grids, Clouds, Syst., Services*, in Lecture Notes in Computer Science, vol. 11819, 2019, pp. 127–139.

[94] Z. Dehghani, "How to move beyond a monothilitic data lake to a distributed data mesh," Tech. Rep., 2019.

[95] S. Lawrenz, P. Sharma, and A. Rausch, "Blockchain technology as an approach for data marketplaces," in *Proc. Int. Conf. Blockchain Technol.*, Mar. 2019, pp. 55–59.

[96] A. Radhakrishnan and S. Das, "Data markets for smart grids: An introduction," in *Proc. IEEE Innov. Smart Grid Technol.-Asia (ISGT Asia)*, May 2018, pp. 1010–1015.

[97] G. S. Ramachandran, R. Radhakrishnan, and B. Krishnamachari, "Towards a decentralized data marketplace for smart cities," in *Proc. IEEE Int. Smart Cities Conf. (ISC)*, Sep. 2018, pp. 1–8.

[98] A. Ganti. (2019). *Clearinghouse*. [Online]. Available: https://www.investopedia.com/terms/c/clearinghouse.asp

[99] K. R. Ozyilmaz, M. Dogan, and A. Yurdakul, "IDMoB: IoT data marketplace on blockchain," in *Proc. Crypto Valley Conf. Blockchain Technol. (CVCBT)*, Jun. 2018, pp. 11–19.

[100] A. Colman, M. J. M. Chowdhury, and M. Baruwal Chhetri, "Towards a trusted marketplace for wearable data," in *Proc. IEEE 5th Int. Conf. Collaboration Internet Comput. (CIC)*, Dec. 2019, pp. 314–321.

[101] S. W. Driessen, G. Monsieur, W.-J. van den Heuvel, and A. Damian, "Validated data quality assessment with 'skin in the game': A smart contract approach," in *Proc. SummerSOC*, 2021, pp. 119–130.

[102] *The Software Product Quality Model*, Standard ISO25000, International Organization for Standardization, 2011. [Online]. Available: https://iso25000.com/index.php/en/iso-25000-standards/iso-25010

[103] G. Möllering, "The nature of trust: From Georg simmel to a theory of expectation, interpretation and suspension," *Sociology*, vol. 35, no. 2, pp. 403–420, 2001.

[104] M. Frederiksen, "Trust in the face of uncertainty: A qualitative study of intersubjective trust and risk," *Int. Rev. Sociol.*, vol. 24, no. 1, pp. 130–144, Apr. 2014, doi: 10.1080/03906701.2014.894335.

[105] M. L. Kersten, S. Idreos, S. Manegold, and E. Liarou, "The researcher's guide to the data deluge: Querying a scientific database in just a few seconds," *Proc. VLDB Endowment*, vol. 4, no. 12, pp. 1474–1477, Aug. 2011.

[106] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger, "Design and implementation of a geographic search engine," in *Proc. 7th Int. Conf. Modelling, Identificat. Control (ICMIC)*, 2015, pp. 1–6.

[107] S. Bajoudah, C. Dong, and P. Missier, "Toward a decentralized, trust-less marketplace for brokered IoT data trading using blockchain," in *Proc. IEEE Int. Conf. Blockchain (Blockchain)*, Jul. 2019, pp. 339–346.

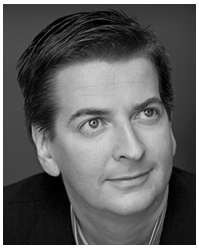
[108] C. Stedman and J. Vaughan. *What is Data Governance and Why Does it Matter?* Accessed: Nov. 19, 2021. [Online]. Available: https://searchdatamanagement.techtarget.com/definition/data-governance

[109] D. Gianni, "Data policy definition and verification for system of systems governance," in *Modeling and Simulation Support for System of Systems Engineering Applications*. Hoboken, NJ, USA: Wiley, 2014, ch. 5.
[110] K. Irion, "Government cloud computing and national data sovereignty," *Policy Internet*, vol. 4, nos. 3–4, pp. 40–71, Dec. 2012.
[111] A. Munoz-Arcentales, S. López-Pernas, A. Pozo, Á. Alonso, J. Salvachúa, and G. Huecas, "An architecture for providing data usage and access control in data sharing ecosystems," *Proc. Comput. Sci.*, vol. 160, pp. 590–597, Jan. 2019.
[112] A. F. Westin, "Privacy and freedom," *Administ. Law Rev.*, vol. 22, no. 1, pp. 101–106, 1969.
[113] B. Levine. (2018). *Nebula Genomics Readies a Marketplace to Sell a Precious Dataset: You*. [Online]. Available: https://martech.org/nebula-genomics-readies-a-marketplace-to-sell-a-precious-dataset-you/
[114] G. Zyskind, O. Nathan, and A. Pentland, "Decentralizing privacy: Using blockchain to protect personal data," in *Proc. IEEE Secur. Privacy Workshops (SPW)*, May 2015, pp. 180–184.
[115] S. De Capitani di Vimercati, S. Foresti, G. Livraga, and P. Samarati, "Toward owners' control in digital data markets," *IEEE Syst. J.*, vol. 15, no. 1, pp. 1299–1306, Mar. 2021.
[116] P. Missier, S. Bajoudah, A. Capossele, A. Gaglione, and M. Nati, "Mind my value: A decentralized infrastructure for fair and trusted IoT data trading," in *Proc. ACM Int. Conf. Ser.*, Oct. 2017, pp. 1–8.
[117] P. Koutroumpis, A. Leiponen, and L. D. Thomas, "The (unfulfilled) potential of data marketplaces," ETLA Work. Papers 53, 2017.
[118] C. Niu, Z. Zheng, F. Wu, X. Gao, and G. Chen, "Trading data in good faith: Integrating truthfulness and privacy preservation in data markets," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, Apr. 2017, pp. 223–226.
[119] M. Reiter, U. Breitenbücher, S. Dustdar, D. Karastoyanova, F. Leymann, and H.-L. Truong, "A novel framework for monitoring and analyzing quality of data in simulation workflows," in *Proc. IEEE 7th Int. Conf. eSci.*, Dec. 2011, pp. 105–112.
[120] M. D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, pp. 1–9, Mar. 2016.
[121] W. Zou, D. Lo, P. S. Kochhar, X.-B.-D. Le, X. Xia, Y. Feng, Z. Chen, and B. Xu, "Smart contract development: Challenges and opportunities," *IEEE Trans. Softw. Eng.*, vol. 47, no. 10, pp. 2084–2106, Oct. 2021.
[122] P. G. Neumann, "Risks of automation: A cautionary total-system perspective of our cyberfuture," *Commun. ACM*, vol. 59, no. 10, pp. 26–30, 2016.
[123] H. Yoo and N. Ko, "Blockchain based data marketplace system," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2020, pp. 1255–1257.
[124] D. Nasonov, A. A. Visheratin, and A. Boukhanovsky, "Blockchain-based transaction integrity in distributed big data marketplace," in *Proc. Int. Conf. Comput. Sci.*, in Lecture Notes in Computer Science, vol. 10860. Cham, Switzerland: Springer, 2018, pp. 569–577, doi: 10.1007/978-3-319-93698-7_43.
[125] C. Banton. (2021). *Escrow Definition: Types, Examples, Pros & Cons*. [Online]. Available: https://www.investopedia.com/terms/e/escrow.asp
[126] S. Sahoo and R. Halder, "Traceability and ownership claim of data on big data marketplace using blockchain technology," *J. Inf. Telecommun.*, vol. 5, no. 1, pp. 1–27, 2020, doi: 10.1080/24751839.2020.1819634.
[127] D. López and B. Farooq, "A multi-layered blockchain framework for smart mobility data-markets," *Transp. Res. C, Emerg. Technol.*, vol. 111, no. Jan. 2019, pp. 588–615, 2020, doi: 10.1016/j.trc.2020.01.002.
[128] P. Gupta, V. Dedeoglu, K. Najeebullah, S. S. Kanhere, and R. Jurdak, "Energy-aware demand selection and allocation for real-time IoT data trading," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Sep. 2020, pp. 138–147.
[129] J. Cáceres, L. M. Vaquero, L. Rodero-Merino, Á. Polo, and J. J. Hierro, *Service Scalability Over the Cloud*. Boston, MA, USA: Springer, 2010, pp. 357–377, doi: 10.1007/978-1-4419-6524-0_15.
[130] Y. Zhao, Y. Yu, Y. Li, G. Han, and X. Du, "Machine learning based privacy-preserving fair data trading in big data market," *Inf. Sci.*, vol. 478, pp. 449–460, Apr. 2019, doi: 10.1016/j.ins.2018.11.028.
[131] T. Hayashi and Y. Ohsawa, "TEEDA: An interactive platform for matching data providers and users in the data marketplace," *Information*, vol. 11, no. 4, p. 218, Apr. 2020.
[132] S. R. Niya, B. Jeffrey, and B. Stiller, "KYoT: Self-sovereign IoT identification with a physically unclonable function," in *Proc. IEEE 45th Conf. Local Comput. Netw. (LCN)*, Nov. 2020, pp. 485–490.
[133] H. T. T. Truong, M. Almeida, G. Karame, and C. Soriente, "Towards secure and decentralized sharing of IoT data," in *Proc. IEEE Int. Conf. Blockchain (Blockchain)*, Jul. 2019, pp. 176–183.
[134] R. J. C. J. Cyganiak and B. McBride, "Resource description framework (RDF)," World Wide Web Consortium, Cambridge, MA, USA, Tech. Rep., 2014. [Online]. Available: https://www.w3.org/TR/rdf11-concepts/
[135] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, "Query-Market demonstration: Pricing for online data markets," *Proc. VLDB Endowment*, vol. 5, no. 12, pp. 1962–1965, Aug. 2012.
[136] T. Hayashi and Y. Ohsawa, "The acceptability of tools for the data marketplace among firms using market research online communities," *Proc. Comput. Sci.*, vol. 176, pp. 1613–1620, Jan. 2020, doi: 10.1016/j.procs.2020.09.184.
[137] D. M. L. Martins, G. Vossen, and F. B. De Lima Neto, "Intelligent decision support for data purchase," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Aug. 2017, pp. 396–402.
[138] D. M. L. Martins, G. Vossen, and M. Maleszka, "Supporting online data purchase by preference recommendation," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 3703–3708.
[139] F. Ricci, L. Rokach, and B. Shapira, *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds., 2nd ed. Boston, MA, USA: Springer, 2015. [Online]. Available: https://link.springer.com/book/10.1007%2F978-1-4899-7637-6#about
[140] P. Banerjee, A. Muthaiah, and S. Ruj, "Blockchain enabled data trading with user consent," in *Proc. Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2021, pp. 263–271.
[141] P. Gupta, V. Dedeoglu, S. S. Kanhere, and R. Jurdak, "Towards a blockchain powered IoT data marketplace," in *Proc. Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2021, pp. 3–5.
[142] Y. Na, Y. Joo, H. Lee, X. Zhao, K. K. Sajan, G. Ramachandran, and B. Krishnamachari, "Enhancing the reliability of IoT data marketplaces through security validation of IoT devices," in *Proc. 16th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, May 2020, pp. 265–272.
[143] I. MacInnes, Y. Li, and W. Yurcik, "Reputation and dispute in eBay transactions," *Int. J. Electron. Commerce*, vol. 10, no. 1, pp. 27–54, Oct. 2005.
[144] N. Amblee and T. Bui, "The impact of electronic-word-of-mouth on digital microproducts: An empirical investigation of Amazon shorts," in *Proc. 15th Eur. Conf. Inf. Syst. (ECIS)*, 2007, pp. 36–47.
[145] M. Ha, S. Kwon, Y. J. Lee, Y. Shim, and J. Kim, "Where WTS meets WTB: A blockchain-based marketplace for digital me to trade users' private data," *Pervas. Mobile Comput.*, vol. 59, Oct. 2019, Art. no. 101078, doi: 10.1016/j.pmcj.2019.101078.
[146] D.-D. Nguyen and M. I. Ali, "Enabling on-demand decentralized IoT collectability marketplace using blockchain and crowdsensing," in *Proc. Global IoT Summit (GIoTS)*, Jun. 2019, pp. 1–6.
[147] W. Agahari, "Platformization of data sharing multi-party computation (MPC) as control mechanism and its effect on firms' participation in data sharing via data marketplaces," in *Proc. 33nd Bled eConf.*, 2020, pp. 691–704.
[148] K. V. and T. K. P., "Protecting privacy when disclosing information: K anonymity and its enforcement through suppression," *Int. J. Comput. Algorithm*, vol. 1, no. 1, pp. 19–22, 2012.
[149] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr. Conf.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 3876, 2006, pp. 265–284.
[150] Y. Hu, C. Li, A. Hu, A. Hu, and J. Zhao, "Trading off data resource availability and privacy preservation in multi-layer network transaction," *Phys. Commun.*, vol. 46, Jun. 2021, Art. no. 101317, doi: 10.1016/j.phycom.2021.101317.
[151] R. Miranda, M. L. Pardal, and A. Grilo, "Sensmart: Sensor data market for the Internet of Things," in *Proc. 35th Annu. ACM Symp. Appl. Comput.*, Mar. 2020, pp. 739–746.
[152] X. Ren, P. London, J. Ziani, and A. Wierman, "Datum: Managing data purchasing and data placement in a geo-distributed data market," *IEEE/ACM Trans. Netw.*, vol. 26, no. 2, pp. 893–905, Apr. 2018.

[153] S. Driessen, D. Di Nucci, G. Monsieur, and W.-J. van den Heuvel, "Automated test-case generation for solidity smart contracts: The AGSolT approach and its evaluation," 2021, pp. 1–15, *arXiv:2102.08864*.
[154] M. C. Nguyen and H. S. Won, "Advanced multitenant Hadoop in smart open data platform," in *Proc. ACM Int. Conf. Proc. Ser.*, 2017, pp. 48–51.
[155] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, "A survey on big data market: Pricing, trading and protection," *IEEE Access*, vol. 6, pp. 15132–15154, 2018.
[156] R. Kohavi, N. J. Rothleder, and E. Simoudis, "Emerging trends in business analytics," *Commun. ACM*, vol. 45, no. 8, pp. 45–48, Aug. 2002.
[157] M. Kirkpatrick. (2008). *GNIP: Grand Central Station on the Social Web*. [Online]. Available: https://readwrite.com/gnip_grand_central_station/

STEFAN W. DRIESSEN (Graduate Student Member, IEEE) received the B.S. degrees in mathematics and physics & astronomy from Radboud University, in 2017, and the M.S. degree in data science & entrepreneurship from the Jheronimus Academy of Data Science, 's-Hertogenbosch, in 2019, where he is currently pursuing the Ph.D. degree. His research interests include applying design science ideas to blockchain technology, smart contracts, data markets, and data products.

GEERT MONSIEUR received the Ph.D. degree in computer science from KU Leuven. He is currently an Assistant Professor of data engineering at the JADS and a Visiting Professor at KU Leuven. His research interests include distributed computing, process and data modeling, smart (micro-)service engineering, architectural design patterns, and model-driven development.



WILLEM-JAN VAN DEN HEUVEL is currently a Full Professor of information systems, data governance and data analytics at the Jheronimus Academy of Data Science and Tilburg University, The Netherlands. He is also the Managing Director of the European Research Institute of Services Science (ERISS). His research interests include cross-junction of software service systems and business process management with an emphasis on (global) networked enterprises.