A Survey of Data Marketplaces and Their Business Models

Extended report [Please cite the ACM SIGMOD version of the paper]

Santiago Andrés Azcoitia
IMDEA Networks Institute
Universidad Carlos III Madrid
Leganés, Madrid, Spain

Nikolaos Laoutaris
IMDEA Networks Institute
Leganés, Madrid, Spain

ABSTRACT

Data is becoming an indispensable production factor for the modern economy, matching or exceeding in importance traditional factors such as land, infrastructure, labor and capital. As part of this, a wide range of applications in different sectors require huge amounts of information to feed machine learning models and algorithms responsible for critical roles in production chains and business processes. A variety of data trading entities including, but not limited to data marketplaces, have thus appeared in order to satisfy and match the offer with the demand for data. In this paper, we present the results and conclusions from a comprehensive survey covering 190 commercial data trading entities, the types of data that their trade, as well as their business models and the technologies that they rely upon. We also point to promising open research questions in the areas of data marketplace federation, pricing, and data ownership protection that could benefit the growing ecosystem of data trading entities that we have surveyed.

KEYWORDS

Data Marketplace, PIMS, Data Economy, Business Model

ACM Reference Format:
Santiago Andrés Azcoitia and Nikolaos Laoutaris, 2022, A Survey of Data Marketplaces and their Business Models. *ACM SIGMOD Record, 51(3), Sep'22*, ACM, New York, NY, USA.

1 INTRODUCTION

Paying for information is not a new idea: insiders have been hired and spies have been trained to achieve a competitive advantage while doing business or fighting wars since ancient times. Such primitive information exchanges exclusively involved humans, yet they were often decisive and undeniably influenced the course of history (e.g., Ephialtes betrayal in the Battle of Thermopilae).

Later, with the advent of telecommunications, information was no longer transmitted by people but by electromagnetic signals, and the exchange of information became instantaneous. Later still, computing, electronics and digital communications gave birth to a new generation of sensors and increasingly automated data collection. As a result, the majority of information now flows from machines to humans.

An even more revolutionary twist will likely drive the future growth of the so-called knowledge economy thanks to the internet of things (IoT), artificial intelligence (AI), and ubiquitous communication systems such as 5G. According to IDC, 30% of data will be generated by sensors in real time by 2025 [61]. In the current context of the major digitalisation of the economy, a myriad of applications and data-hungry machine learning (ML) models are - to give a couple of meaningful examples - helping companies and public institutions improve their efficiency, as well as assisting individuals in health issues. This means that machines will join humans as the main data consumers. In some settings, such M2M data exchanges will be required to happen in real time, too.

As digitalisation progresses, machines are increasingly playing a leading role in data value chains that begin with the collection of data through probes and sensors and terminate with their "consumption" by machine learning (ML) models involved in services provided to end users. In fact, the global amount of data created annually is expected to grow by 530% from 2018 to 2025, of which at least 30% will come from machine-to-machine communications [61]. This influx of new and existing data is thus accelerating the data economy, which is estimated to reach US$2.5 trillion globally by 2025 [38]. Regardless of whether "data" is a commodity like oil, capital, an asset, or similar to labor [6], it is undoubtedly becoming a cornerstone of modern economic systems.

A number of open directories exist on the web for listing data trading entities (DTE) [21, 25, 60]. Such directories loosely tag as '*data marketplace*' heterogeneous entities with hugely different objectives, focus, customers, and business models, etc. For example, traditional *data providers* have been collecting, enriching, and curating public and private information from different sources and silos for years, building successful business models mainly in the areas of marketing (Acxiom, Experian, etc), financial, and business intelligence (Bloomberg, Thomson Reuters, etc.). More recently, data marketplaces (DMs), i.e., two-sided platforms for matching data sellers and data buyers and mediating in data exchanges and transactions, have also arrived on the scene [64].

First-generation *general-purpose DMs* are being complemented by *niche DMs* that target specific industries (e.g., Caruso for the connected car, Veracity for energy and transportation), and cover data sourcing for specific innovative purposes, such as feeding ML algorithms (e.g., Mechanical Turk, DefinedCrowd), or trading IoT real-time sensor data (e.g., IOTA, Terbine). Additionally, some leading *data-management systems* (e.g., Snowflake, Cognite) and *niche* digital solutions (e.g., Carto, Openprise, LiveRamp) are integrating secure data exchange features and capabilities to their existing products with the aim of breaking data silos [29].

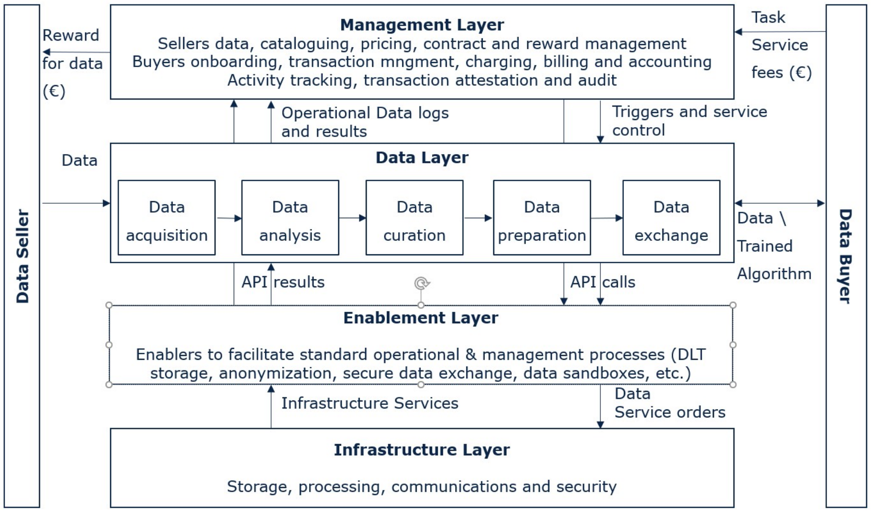

Data Seller
Reward for data (€)
Data
Management Layer
Sellers data, cataloguing, pricing, contract and reward management
Buyers onboarding, transaction mngment, charging, billing and accounting
Activity tracking, transaction attestation and audit
Task
Service fees (€)
Operational Data logs and results
Triggers and service control
Data Layer
Data acquisition
Data analysis
Data curation
Data preparation
Data exchange
Data \ Trained Algorithm
Data Buyer
API results
API calls
Enablement Layer
Enablers to facilitate standard operational & management processes (DLT storage, anonymization, secure data exchange, data sandboxes, etc.)
Infrastructure Services
Data Service orders
Infrastructure Layer
Storage, processing, communications and security

Figure 1: A layered approach to data trading

Aided by recent legislative developments, including the General Data Protection Regulation in the EU or the California Consumer Privacy Act [26, 52], *Personal Information Management Systems (PIMS)* have appeared with the purpose to empower individuals to take back control of their personal information currently being collected by Internet service providers, with little or no consent. Some of them have also implemented marketplaces for helping users monetize their personal data [67].

Purpose of this paper: The complex ecosystem of DTEs combined with the inherently complex nature of data as an asset, which is freely replicable and non-perishable, serves a wide range of uses, and holds an inherently combinatorial and aprioristically unknown value that depends on the buyer and the use case [1, 51], limits our ability to understand this evolving ecosystem and to contribute, from the research point of view, to its evolution. Therefore, in this paper we strive to present a comprehensive survey covering commercial data trading entities, and the wide spectrum of data types, business models, and technologies that they encompass. We have developed a taxonomy system summarized in Tab. 2 for the more than 100 DTEs listed in Tab. 4. This analysis points to a number of open challenges that we believe should attract the attention of the research community, including issues of pricing, federation of different marketplaces, and data ownership protection.

The remainder of the paper is structured as follows:

- Section 2 introduces the data value chain and the concept of '*business model*'.
- Section 3 presents the scope of the survey, and characterizes a catalog of business models that we compiled during our study.
- Section 4 provides more detail on how different entities trade and exchange data on the Internet.
- Section 5 presents state-of-the-art novel related proposals coming from the research community.
- Section 6 summarizes a series of key challenges that we identified while surveying the area and points to promising research directions.
- Finally, Sect. 7 summarizes the key takeaways from our analysis and presents some further trends in this fast changing ecosystem.

As a starting point, we provide some background, and we introduce some of the terms and definitions we will use throughout the paper, along with our view of the data value chain.

2 UNDERSTANDING THE DATA TRADING VALUE CHAIN

In the context of data trading, *actors* in the value chain are legal entities or individuals playing an effective role in producing any data-driven service or data product, be it intermediate or final, that is offered and eventually acquired or exchanged in the market. We will generally refer to them as data trading entities or DTEs. Our survey aims to understand what the roles of such DTEs are, how they interact with other DTEs, how they do business, and what mechanisms they use to set prices for data. We encapsulate all this information in the concept of a *business model*, a term that has been defined in various ways in the literature [56]. For the purpose of this paper, we will refer to a DTE's *business model* as the description of its value proposition within the chain, the processes or activities it covers, the inputs it requires, and the outputs it provides the market with, as well as the relationship the entity maintains with other actors [18].

Understanding the data value chain is a key first step in order to identify relevant business models. Previous studies have already explained the data value chain in general [19, 36], and specific contexts [49, 50]. From a broad data trading perspective, Fig. 1 shows a diagram of four stacked functional layers that allow sellers and buyers to connect. We will later use this to position and classify actors in the market.

At the bottom, the *infrastructure layer* provides the basic processing, secure storage and communication functions to the upper layers in the stack.

On top of such infrastructure, the *enablement layer* provides generic application programming interfaces (APIs) and functions to DTEs. Some solutions and PDKs in the market do not intend to directly provide services to the end users, but rather to provide a platform with common useful functions that *enable* other DTEs to carry out a controlled data exchange which optionally may involve an economic transaction.

In the next level, a more technical and operational *data layer* deals with data processing itself and responds to the effective delivery of data or data-driven services to end-customers, be they consumers, or other DTEs. Reaching from data collection or extraction to its final delivery to the end consumer, this process usually requires intermediate preprocessing, curation and data enrichment steps. In addition, it may involve third parties whose data is acquired and combined, and therefore other secure data exchanges.

Finally, the top *management layer* deals with data discovery, coordinates transactions, keeps track of contracts and service level agreements, and ensures the accountability and transparency of all the operations and processes in the data layer. In contrast to the operational *data layer* immediately below, it works with metadata and transactional data. Other functions of the management layer include helping data-owners catalogue, structure and price their data offer, governing data transactions (e.g., through contract management, charging, billing and accounting processes), and increasing the overall transparency of data trading. In the case of transactions involving data from multiple sellers, it is also in charge of distributing the resulting payments among them.

Note that our definition allows for cascading transactions, which is oftentimes the case before sufficiently processed data is transformed to a data-driven service to end-consumers. For example, a model that outputs consumer segmentation data at postcode level requires at least the following steps: i) gathering anonymized segmentation data (often from disparate sources), ii) combining such information with geo-located identity data into a single coherent dataset, and iii) aggregating this output into individual postcodes by processing it together with postcode border shapefiles (often obtained from a third party, too) in a geographical information system.

3 A TAXONOMY OF BUSINESS MODELS

3.1 The ecosystem of data trading

We initially checked more than 190 companies offering data products on the Internet. After a brief initial review, we selected 104 of them to analyze in detail (final list available in Table 4 in the appendix). We discarded concept projects, online advertising platforms, and Internet service providers not specifically offering data products.

The final set includes companies of different sizes from 23 countries, as Fig. 2 shows. We collected information published by these companies on their web-sites to better understand their business models. For example, we investigated the data that they trade, how they collect and manage it, whom they sell it to, exactly what they provide to customers, and how they deliver and price their services. Furthermore, we collected information about when these companies were founded (half of them in 2016 or later) and how many employees they have (40% of them have fewer than 20 employees).

Most companies in our sample are either *scaling* their customer base (29) or are in *commercial* development stage (61). In addition, we have included *developing* companies working in new innovative concepts around IoT, personal and ML data, or integrating blockchain in decentralized architectures (e.g., DataBroker/Settlemint and Dataeum). Finally, we chose not to include any *open data* providers and repositories, but instead focus only on commercial entities offering paid data products.

Appendix A thoroughly explains the methodology we followed, including the set of questions we set out to answer, how we gathered and analyzed the information, and some limitations of our study. In addition, appendix B shows the list of the entities included in the scope of this survey.

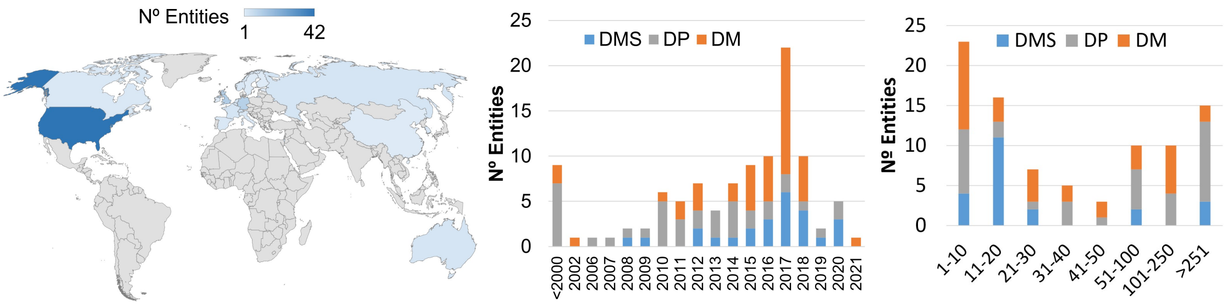
3.2 Data trading entities by customer segment

First, we found that the business models of DTEs heavily depend on who they consider their customers to be, which in turn depends on which side of the chain they approach data trading from. *Data management systems* (DMSs) focus on managing the information an enterprise or individual owns. Conversely, traditional *data providers* (DPs) focus on data consumers, and conceal data owners and often even their partners when selling their products. Whereas the former approached data trading in order to allow secure data exchanges within an organization or to authorize third parties, the latter implemented data trading platforms to complement their existing products or services with those of third parties. In addition, *data marketplaces* (DMs) were conceived from the beginning as two-sided platforms dealing both with buyers and sellers.

Within the scope of the survey, we included 41 DMs and 25 DMSs. As far as DPs are concerned, they often provide their products in commercial DMs, and we managed to list 2,015 of them selling their products in a sample of nine public or semi-private DMs. Hence, they are by far the most frequent business model within DTEs. Since the way they operate is often similar, we took a diversified sample of 38 to understand how they deliver data and how they price their products.

3.3 Data trading business models

In this section we dive deeper into the different business models and their variations, which we summarize in Tab. 1. Interestingly, we find that some PIMSs and DMs only implement partial data trading functionality. Such *enablers* provide a range of solutions that includes, for example, anonymizing personal information (AirCloak), providing an homogeneous anonymized identity to buyers (Datavant), facilitating secure exchanges (Cybernetica), or empowering individuals to exert their rights on the information that data providers hold about



(a) Entities by country **(b) Entities by year of foundation** **(c) Entities by nº employees**

Figure 2: Summary of entities included in the survey

Table 1: Taxonomy of data trading business models

	Data Providers (DP)	**Data Marketplaces (DM)**	**Data Management Systems (DMS)**
End-to-end DTEs	Service Providers	General-purpose DM	Embedded DM
	Data Providers	Niche-DM	PIMS
	Private marketplaces (PMP)		Survey PIMS
Enablers		DM enablers (DME)	PIMS-enabler

them (Saymine). When it comes to charging and billing, enablers usually charge for transactions (e.g., calls to the API, volume of data processed, etc.). Even though some enablers focus on specific types of data (e.g., IoT-related, data for ML models, personal data), or industries (e.g., health), we do spot some *general-purpose* enablers as well (e.g., those providing secure data exchange of distributed data).

With regards to entities providing full-fledged end-to-end seller-to-buyer solutions, Tab. 2 summarizes their main characteristics (in rows, along with the section the topic is dealt with in the paper) and the differences between their business models we identified (in columns), which we further explain in the next sub-sections.

3.3.1 Providers. We consider two types. *Data Providers* (DPs, aka vendors [64]) are entities which provide *data* as a product, be they raw or enriched data, access to information through a graphical user interface, or information contained in reports to third parties. They usually combine data from different sources (e.g., from the public Internet, from partners, or from other providers) to enrich their products and add value to their offer.

Service Providers (SPs) are entities providing digital services to end-customers, be they individuals or enterprises, based on data they own, or on that which they collect from the Internet, or acquire from third parties. Examples of them are Clearview.ai, a company that provides identity data based on pictures of people publicly available on the Internet, or Factual, which offer marketing insights based on the movement of people. The boundaries between data and service providers are often blurry: are not personal identifications provided by Clearview.ai or insights by Factual data in the end?

From our point of view, supply side platforms and demand side platforms are SPs in the online marketing industry. The former allow publishers and digital media owners to manage and sell their ad spaces, whereas the latter allow advertisers to buy such advertising space, often by means of real-time automated auctions. Also in online marketing, data management platforms (DMPs) refer to audience data management systems that allow advertisers to enrich their audience data with that provided by the DMP. Some marketing-related SPs (Liveramp, Lotame, Openprise, among others) are integrating *private marketplaces* (*PMPs*) into their platforms to allow secure exchanges, monetization, trading and integration of audience data from trusted partners (among them the so-called *data brokers*) within the platform. Such marketplaces are frequently an add-on to DMP subscriptions, and therefore can only be accessed by their users.

Despite the fact that the term *PMP* often refers to data trading platforms operated by marketing-related service providers, similar business models have also flourished in trading geolocated data (Carto, Here), business technographic data (Crunchbase), and financial data (Factset, Quandl, Refinitive). They all provide their users with a marketplace to enrich their data with relevant second-party and third-party data. As opposed to public or semi-private DMs, data exchange in PMPs is a *private functionality* of data and service providers that complements their main value proposition, and hence is only accessible by their customers on the buy side, or authorized data partners on the sell side.

Interestingly, as well as directly commercializing their services through their websites, DPs and SPs also use intermediaries to advertise their services, provide access to free samples of data, or offer specific data products. We found that 45% of data brokers (like Experian, Acxiom or Gravy Analytics) that offer their products through marketing-related PMPs (e.g., Liveramp, TheTradeDesk or LOTAME) commercialize their products in other DMs such as AWS or DataRade, too. This is also the case with providers such as RepRisk, Equifax or Arabesque S-Ray, which make use of financial PMPs.

Table 2: Summary of business models

	Data Trading Entities (DTE)					
	Providers		**Data Marketplaces (DM)**		**Data Management Systems (DMSs)**	
Concept (Sect.)	**DP/SP**	**PMP**	***General-purpose* DM**	***Niche* DM**	***Embedded* DM**	**PIMS**
Data exchange (3.3)	Public, semi-private, private	Private	Public / Semi-private		Private	Public / Semi-private
Scope (4.1)	Focused		Diversified	Focused		
Type of data (4.1)	Any	Specific data to be used within their service / platform	Any	Industry or type-specific	Data exchanged within the system	Personal data
Roles / Players interacting (3.3)	Partners, Customers		Sellers, buyers		Owner, requester	Users, data Providers, buyers
Gets data from (3.3)	Internet, self-generated, partners, users	Partners, Data providers	Data providers	Data providers, self-enriched	Data providers	Users, Data providers
Provides buyers with (3.3)	API, datasets	API, access to data through the system	API, datasets		API, Access to data through the system	API, Key to decrypt data
Owners get access through (3.3)	Partnership	Partnership, the service platform	Web-services		Data Management platform	Mobile App Web services
Buyers get data through (3.3)	Web-services, APIs	Web-service, the service platform	Web-services	Web-services, APIs	Data Management platform	Web-services, APIs, compatible systems
Access pricing for buyers (3.3)	Subscription, pay for data	Included in the main platform	Predominantly free. Some freemium, subscription and data delivery charges		Add-on to the data management Platform	Pay for data
Access pricing for sellers (3.3)	Partnership (when applicable)	Partnership, time subscription	Predominantly free. Some freemium subscription, and revenue-share charges		Subscription to the platform	Free
Data pricing schemes (4.3)	Fixed one-off, subscription, customized, volume-based	Subscription, domain-specific (e.g., cost per click, cost per 1,000 impressions, ...)	Fixed one-off, subscription and customized	Customized, volume/usage-based, fixed one-off	Open	Open, bid by buyer
Data price set by (4.4)	Platform	Platform, buyers	Platform, providers		Open	Users, Platform
Payment (4.5)	Fiat currency			Fiat currency, token	Open	Token, fiat currency
Platform type (4.8)	Centralized		Centralized or decentralized		Centralized	Decentralized

3.3.2 Data Marketplaces.

DMs are mediation platforms that put providers in touch with potential buyers, and manage data exchanges between them. Such exchanges usually involve some kind of economic transaction, as well, either through payments in fiat currency or in a cryptocurrency often created and controlled by the platform. DMs are either public - i.e., open to any data seller or buyer - or semi-private, meaning any seller or buyer is subject to the approval of the platform in order to be allowed to trade data. Furthermore, DMs often deal with data categorization, curation and management of metadata to help buyers discover relevant data products.

Whereas *general-purpose* DMs like AWS, Advaneo or DataRade trade any type of data, *niche* DMs are focused on certain industries (martech, automotive, energy) and on certain types of data (spatio-temporal data, or that coming from IoT sensors). By analyzing the date of foundation, we spotted a clear trend towards real-time data streaming marketplaces to harness the potential of IoT (e.g., IOTA, Terbine), and those specialized in training ML models (e.g., Skychain, Ocean Protocol), both very active lines of scientific research (see Sect. 5).

The large number of identified marketplaces, each one having proprietary on-boarding processes, access protocols/APIs and user-interfaces, makes it challenging for data providers to establish presence in all of them and thus reach the widest possible audience. The fragmentation of the DM ecosystem calls for establishing inter-operability standards that will allow different marketplaces to federate as discussed in Sect. 6 (*Challenge 1*). Sellers and buyers are often invited to subscribe for free to the platform. However, some platforms charge for freemium subscriptions or charge IaaS-like fees for delivering data. A few of them opt for charging sellers according to the money they make through the platform, either through commissions or revenue sharing.

In addition, buyers oftentimes pay marketplaces for data. Both the data seller and the platform are in charge of setting the prices for data products - in most cases one-off charges for downloading or gaining access to datasets, or periodic subscriptions to data feeds in *general-purpose* DMs. Conversely, *niche* DMs more frequently resort to volume or usage-based charging for APIs, and price customization depending on who the data buyer is and on what the purpose of purchasing the data is.

Some DMs build on top of *data marketplace enablers* (DMEs). For example, Ocean Protocol provides marketplace functionality for ML data trading, whereas GeoDB and Decentr are DMs that use Ocean Protocol.

3.3.3 Data Management Systems.

On the one hand, enterprise DMSs are increasingly offering add-ons to carry out secure data exchanges within an organization, and to enrich its corporate information base by acquiring data from second or third-party providers. Such *embedded DMs* - meaning they are built in an already existing DMS - rarely include full marketplace functionality, but rather restrict themselves to securing data exchanges, and to controlling the delivery and access to data assets within the walled-garden of information under their control of each customer. Some of them charge IaaS-like fees for delivering data, and a recurring subscription fee to authorized sellers.

On the other hand, *PIMSs* look to empower individuals to take control of their personal data, and act as a single point of control to manage them. They leverage recent data protection laws so as to let users collect personal information controlled by digital service providers, exercise their erasure or modification rights as granted by law, manage permissions of mobile apps to give away their data, manage cookie settings, etc.

In addition, some of them seek their users' consent to share their personal information with third parties through the platform in exchange for a reward. Almost half of the surveyed *PIMSs* include marketplace functionalities, and focus on trading personal data for marketing purposes such as user profiling and ad targeting. Therefore, they leave data subjects (as the owners) and data providers to negotiate fees for consenting to get access to their data. This way they become personal data brokers, letting users monetize their data, and controlling who has access to it and for what purposes.

Recently, health-related *PIMSs* (Longenesis, HealthWizz, MedicalChain) specialize in managing healthcare-related information of their users. We found that health-related *PIMSs* often resort to blockchains to provide additional security to such sensitive data, and comply with a strong sectorial regulation. However, it is unclear whether and how their solutions protect against data replication and distribution off the chain.

Finally, *survey PIMSs* (e.g. Citizen.me, ErnieApp or People.io) aim to facilitate targeted marketing surveys among their users, leveraging information about their profile to offer an accurately targeted audience, and rewarding users for participating in the processes.

As opposed to enterprise DMSs, *PIMSs* are more decentralized platforms that often leverage the users' devices to store information, and they are always offered for free to individuals. Some charge one-off fees, subscription, or data delivery fees to potential data buyers.

4 QUANTITATIVE INSIGHTS INTO DATA TRADING

Having characterized qualitatively the business models of DTEs, this section takes a closer quantitative look into crucial aspects of data trading. In the following sections, we tackle questions related to:

- the kind of data being traded (Sect. 4.1),
- the source and the target of DTEs (Sect. 4.2),
- pricing schemes (Sect. 4.3) and
- responsible parties to set the prices (Sect. 4.4),
- payment methods (Sect. 4.5),
- billable concepts and charging (Sect. 4.6),
- business models used to trade data (Sect. 4.7),
- type of storage and architecture (Sect. 4.8),
- how buyers can test data (Sect. 4.9), and
- general data security issues (Sect. 4.10)

4.1 What kind of data is being traded?

Very different kinds of data are being traded in the market. In fact, DTEs are often classified based on the kind of data they trade. For example, we will talk about *marketing DPs* or *marketing PIMS*, meaning DTEs specialized in providing data or managing and trading personal information for marketing-related purposes. We will also discuss the aforementioned *general-purpose* DTEs aiming to trade any kind of data.

Figure 3a shows a breakdown of the kind of data traded by DMSs (in blue), DMs (in orange) and DPs (in grey). There are notable differences in what kind of data entities do trade depending on their business model.

- DPs specialize in a market *niche*, either a specific type of data or a customer segment. Only one of them (Quexopa) is publicly focusing on collecting and delivering data for a certain region (Latin America). Even though the range of data DPs deal with is diverse, it turns out that most providers in our sample are related to marketing, corporate, contact or financial data.
- Within DMSs, *PIMSs* focus on personal and healthcare-related data, whereas business-oriented DMSs are usually designed to trade different types of corporate data.
- With regards to DMs, at least 14 of them are *general-purpose* and trade *any* kind of data, whereas *niche* DMs deal with healthcare, automotive, IoT-related, trading or alternative investment data.

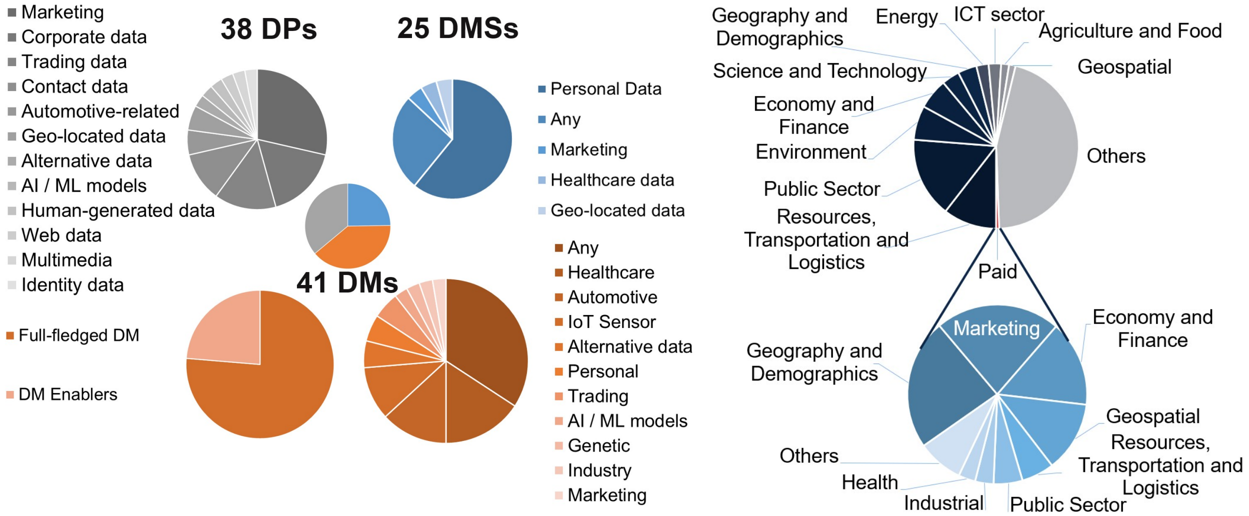
Focusing on *general-purpose* DMs, we carried out a deeper analysis, drilling down to the level of data products, to better understand what are the categories of data most frequently offered in those markets. For that purpose, we gathered public information about almost two million data products from AWS, DIH, Advaneo, DataRade, Knoema, Snowflake, DAWEX, Carto, Veracity, Crunchbase and Refinitiv, and matched their category tags at a high level.

Figure 3b presents the most frequent data categories of data products in *general-purpose* DMs. The pie chart on the left includes free and paid data products, whereas the one on the right includes only those that are paid (10,860 products). '*Marketing*' and '*Economy and Finance*' fall among the most popular categories for paid data products. Moreover, the presence of '*Geography and Demographics*' and '*Geospatial*' data marks the importance of geo-located data in the sample, as well.

Other interesting takeaways from this analysis are that most data products in *general-purpose* DMs are made available for free, and that some of them such as DIH, Advaneo, and Google Cloud Marketplace lack any significant offer of paid products. We observe that these free data products are either open data from public repositories, or samples uploaded by data providers.

Surprising though it may seem in the case of entities whose aim is to make profit, marketplaces like DIH or Advaneo collect and link open data made available by authorities or public institutions. They give up on generating revenues from reselling paid data, and they monetize the effort to organize and facilitate the exploitation of free open data assets in other different ways. For example, some DMs offer free datasets as part of a subscription to the platform (e.g., Carto), whereas others charge for processing and integrating data within the platform (e.g., Advaneo, and those managed by cloud service providers). Such a vast amount of data may also serve as a '*hook*' for sellers and buyers, and as a complement to third-party paid data products.

Moreover, we find that some providers are making use of public marketplaces to upload outdated samples of their products so that buyers can manipulate them and get to know how useful the whole data product would be for their purposes. This practice would indeed be interesting for DMs, provided it



(a) Breakdown of entities by data category and business model

(b) Breakdown of products in *general-purpose* DMs

Figure 3: Data trading entities and the kind of data they trade

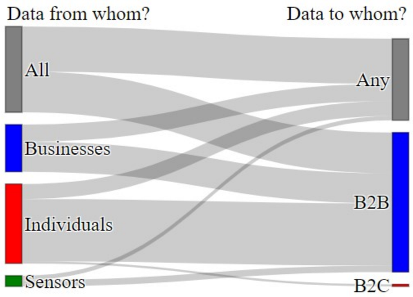


Figure 4: Flows of data in commercial DMs

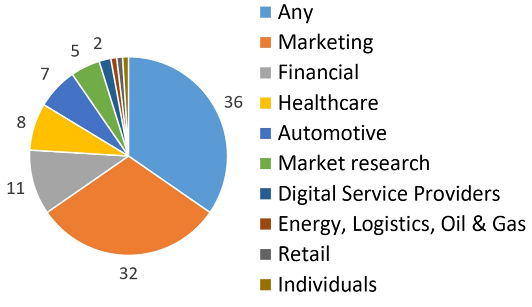


Figure 5: Target of data trading entities

was they who eventually sold the corresponding paid product after the trial. However, data providers usually refer interested buyers to their own commercial channels, and the host marketplace merely acts as a showcase for their products.

4.2 Data from whom? Targeting whom?

Figure 4 shows where or from whom data trading entities get their data from and to whom they intend to sell their data and their services. Data may come from different sources, such as PI owned by individuals, data related to companies, industries, measurements from sensors, etc. However, most data trading entities are designed to trade data from *all* these sources. Even though most entities are clearly oriented to the business market, and we can state that most data nowadays flows to enterprises, no restriction seems to prevent individuals from also acquiring data. DTEs usually target specific industries, and often specific departments within their business customers, which we summarize in Fig. 5. Unsurprisingly, it is marketing and financial departments that DTEs' data and services are most often targeting, with more and more marketplaces oriented to healthcare and to specific industries lately.

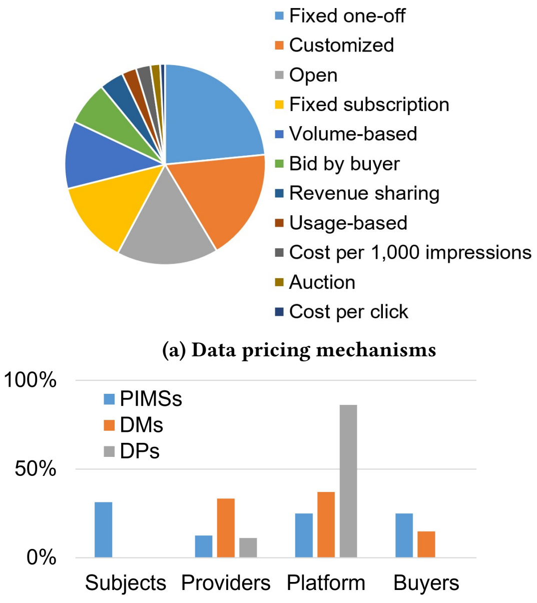
4.3 How is data being priced?

Figure 6a provides a summary of the most widely adopted pricing mechanisms. Most pricing schemes are in line with the current state-of-the-art [58]:

• Most buyers pay a lump-sum as a **fixed one-off** for a dataset, or a **fixed subscription** charge for accessing a stream or a service for a period of time.

• **Customized**. Price (and the product) is personalized depending on who the buyer is and what the data is intended to be used for.

• **Open**. The platform lets buyers and sellers agree on the prices, and eventually consent to the data exchange. This is the preferred approach of *enablers* that focus on facilitating data exchanges but do not involve themselves in setting their economic terms.



(a) Data pricing mechanisms

(b) Stakeholders setting the prices of data

Figure 6: Data transaction pricing

• **Volume-based**. Price directly depends on the volume of information that is downloaded or accessed (e.g., contacts, images), often with volume discounts.
• **Bid by buyer**. Buyers place bids (e.g., in a *PIMS*) that sellers must accept for the transaction to take place. This avoids sellers deciding on an upfront asking price.
• **Usage-based** pricing is frequently used for API calls and offered in tiers. Charges include volume discounts and depend on the number and type of calls.

In addition, we identify some other interesting mechanisms being used in specific contexts.

First, MyDex used to charge transactions using **revenue sharing**: when a buyer purchased the rights to access a user's personal data, the platform claimed its rights to 4% of the revenues that such a buyer made on the platform from that individual. Digi.me, a *PIMS enabler*, also mentions this pricing scheme. Although innovative in terms of pricing data, its feasibility is still to be proven: would a *PIMS* be able to control how much money buyers are making from each user and charge accordingly? None of them are using this scheme now.

Cost per 1,000 impressions, **cost per click** and **percentage of gross media expenses** are specific to PMPs of online advertising platforms (e.g., LiveRamp, Oracle or Kochava), thanks to their end-to-end control of online ad campaigns.

Finally, **auctions** are very popular price setting mechanisms in other fields, and they are widely used in online advertising where advertisers bid in real time to show their ads to a user browsing a certain webpage [57]. Nonetheless, they are not so common when selling data, due to its non-rivalrous nature - meaning selling a piece data to A does not prevent from selling the same data to B, hence A and B do not behave as rivals. Even though some works of research have already defined a whole family of auctions that artificially creates competition among interested buyers [34, 35], we found only one enabler (Ocean Protocol) that mentions auctions as a potential mechanism to set the prices of data products.

4.4 Who sets the price of data products?

The answer to this question again depends on the business model, as Fig, 6b shows. Whereas providers tightly control the price of their data or services, *PIMSs* give more control to their individual users (the actual data subjects), and usually let them agree with buyers on transaction prices. Although DMs usually play an active role in setting the prices for data products on their platform, they always do it in conjunction with providers. In fact, some of them (Dawex, Battlefin) charge for advisory services in setting the prices for their data products. Such advisory services for pricing are empirically provided since developing a more rigorous methodology for data pricing remains an open challenge (*Challenge 2*).

4.5 How do entities deal with payments?

Whereas data providers have traditionally been charged for their services in fiat money (dollars, euros, etc.), 55% of surveyed *PIMSs* and almost 40% of marketplaces are using cryptocurrencies instead. The benefits they seek by using this alternative include an increased speed of transfers, a higher availability if compared to going through banks or establishments, and a greater liquidity. Real-time data exchanges like the ones trading with IoT sensor data are broadly opting for cryptocurrencies when it comes to liquidate payments.

4.6 Do data trading platforms charge users for accessing their services?

DTEs that operate as platforms do not only charge users for the data they consume, but for other concepts such as delivering data, or even just for gaining access to their services. Again, such additional platform charges vary greatly between business models.

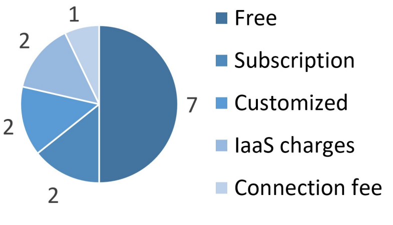
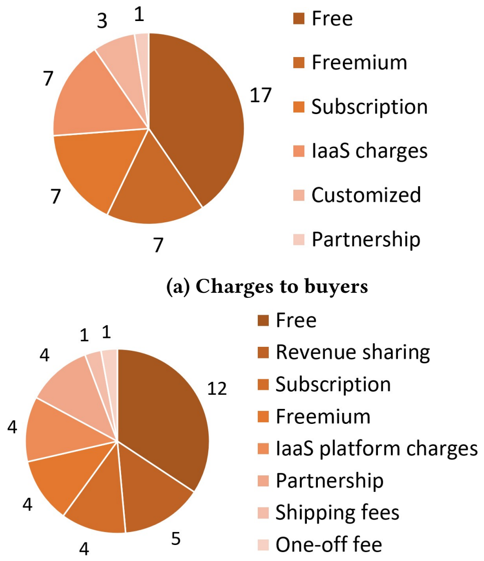


Figure 7: Charges to buyers accessing PIMS

In general, PIMS are free for data subjects. On the one hand, this makes sense since they provide the platform with PI to work with, and also make the promise of increased privacy and data protection more appealing. On the other hand, it raises concern over the profitability of users who are unwilling to share their data and are using PIMS for such purposes. Data



(a) Charges to buyers

(b) Charges to sellers

Figure 8: Charges and pricing in data marketplaces

buyers, who are also usually welcome and free to join the platform, often just pay for the data they acquire. In some cases, potential buyers are asked for a one-off connection fee or charged a periodic subscription (see Figure 7) to get access to the platform. Finally, some PIMS demand details about the buyer signing up to the platform to customize access charges.

Conversely, charging buyers and especially data sellers for access is more usual in the case of DMs (see Figs. 8a and 8b), either through:

- time-based subscriptions, often using a freemium model;
- revenue sharing, where the platform keeps a percentage of the total sales;
- one-off fees to connect to the system.

A few *niche DMs* (Otonomo) and most PMPs offer partnership models to big data sellers, an *ad hoc* agreement to share data frequently used by DPs. Interestingly, a niche DM (Caruso) requires a partnership agreement to be signed by buyers, which requires their participation as shareholders if they are willing to use the platform.

Regarding PIMS and DM-enablers, they welcome full-fledged PIMS to use their technology and usually charge pay-as-you-go IaaS/PaaS-like fees based on the number of API calls or the volume of data they deliver. Some DMs (e.g., AWS or Snowflake) do charge data shipping fees to both parties, too.

4.7 How do entities trade data?

Some specific characteristics of data, in particular its zero-cost replicability and its inherently combinatorial value, make this attractive asset considerably more difficult to be priced and safely traded [58, 65]. In economics, a good or service is called *excludable* if it is possible to prevent consumers who have not paid for it from having access to it. In addition, data is non-depletable and hence a *non-rivalrous* good: selling data to A does not prevent the owner from selling it again to B.

Entities trading data aim at somehow making it *excludable* and therefore a club good. Indeed, this is a key challenge in building a flourishing economy around data. This section provides some additional insights about how entities in our survey are attempting to achieve this goal. In the following subsections, we answer questions regarding where entities take data from, what they provide buyers with, and how users - both from the buy and sell sides - gain access to data.

Since the conclusions are very different for DPs, DMs and PIMS, we present them separately in subsections.

4.7.1 *Data Providers.*

As Fig. 9 shows, DPs leverage the internet and access to exclusive self-enriched data sources to provide buyers with access to data either through APIs or bulk downloads, and preferably through web-services or specific applications. Note that they are not meant to be two-sided platforms, but players oriented to provide their data or their data-driven services to their customers. Should they require proprietary information from third parties, they establish partnerships or bilateral agreements to access such exclusive information, which they eventually enrich and resell. Therefore, DPs control the whole go-to-market process, and conceal the identity of their partners and the sources of their information, unless disclosing them adds any value (e.g., credibility) to their business and hence helps with sales.

As an exception, PMPs integrated in data-driven services (e.g., spatio-temporal data marketplaces integrated in GIS cloud SPs) allow third-party DPs to sell data within their platform. Unlike DMs, PMPs carefully select authorized DPs, who often sign private partnership agreements with them. Moreover, they deliver data to be used within their system or services, and only to their users, which is why such marketplaces qualify as *private*.

4.7.2 *Data Marketplaces.*

Figure 10 shows that DMs collect and sometimes enrich or combine data from different DPs (sellers), who have signed the DM's public terms of use. Similar to DPs, data is often delivered to buyers as a bulk download or through APIs. Although some of them restrict delivery methods to get access to data through their platforms (e.g., AWS marketplace offers access to data stored in Amazon S3 services). they often resort to APIs and web services for users to manage their transactions and data within the system.

4.7.3 *PIMS.*

PIMS collect and manage personal data from individuals. PIMS help users retrieve their PI from third parties like social networks or e-mail services, which they call their *data providers*. Users can share their personal information with the platform, as well , and all their personal data are stored in their devices. Then, PIMS may sell such personal data to potential buyers, or either share it with third parties upon the owners' consent. Most frequently, PIMS users use a mobile application to get access to the service, which also allows them to manage their consent to share their PI and monitor data transactions. Most of them provide buyers with

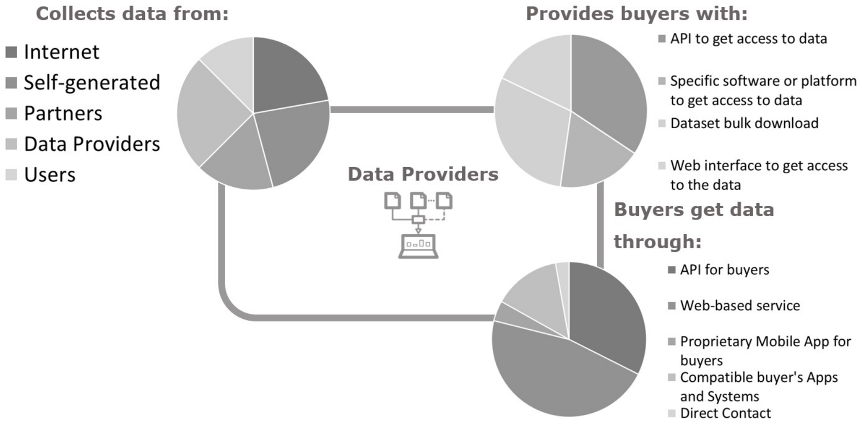


Figure 9: How do data providers work?

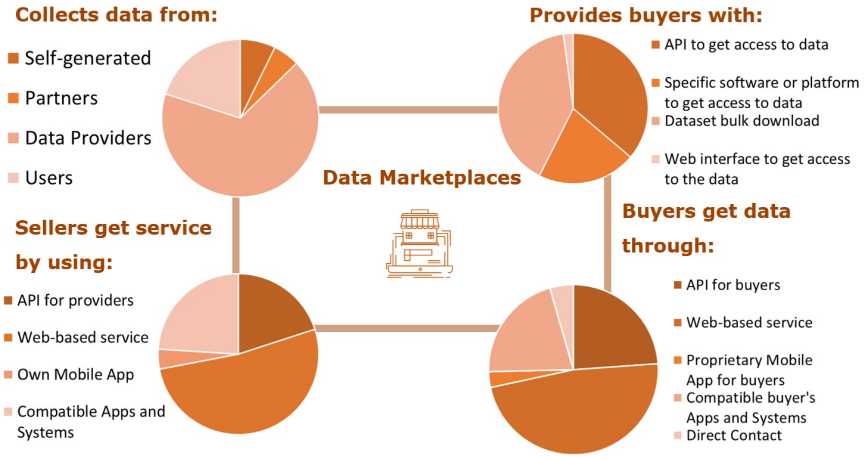


Figure 10: How do surveyed data marketplaces work?

APIs or web services to gain access to data. As opposed to DPs and DMs, some PIMS ask entities willing to gain access to acquired data to integrate their apps and systems (MyDex, GeoDB, DataWallet).

PIMS deliver data in technologically innovative ways. For example, some of them provide access to encrypted data streams by sending temporary keys that are revoked once the subscription expires. Some of them resort to hashed temporary URLs to provide buyers with access to data for a certain period.

Some PIMS still do not provide an automated platform for buyers to get data and results, but instead they negotiate directly with buyers, and generate the data to be shared with data buyers case by case.

4.8 How do entities store data products?

PIMSs usually opt for a more decentralized architecture by leveraging data subjects or providers to store and process users' data. With some exceptions, they avoid making copies of personal information, which is retrieved from the users' personal data storage, instead. On the contrary, DPs and DMs have traditionally preferred a centralized architecture.

Figure 12 shows a trend towards decentralization of the storage of both data products and transactional data. In fact, distributed ledger technologies (DLTs) are increasingly being used to store transactional or management data related to data trading. Due to the high cost of storing data in a DLT, it is not yet being considered as a alternative for storing data for sale, except for specific concept models and developing prototypes related to healthcare (BurstIQ, MedicalChain), marketing (Datum, already closed) and automotive (AMO), whose feasibility is yet to be proven.

4.9 Can buyers "try" before buying?

Buyers cannot be sure about the true value of a dataset before they get access to it and, e.g., train a ML algorithm and test its resulting accuracy/performance. This chicken-and-egg problem, known as Arrow's information (or disclosure) paradox, often deters potential buyers from actually purchasing data. A big challenge for the data economy is thus to come up with ways to reduce the uncertainty for buyer [67]. Next we look at how surveyed entities approach this challenge (*Challenge 3*).

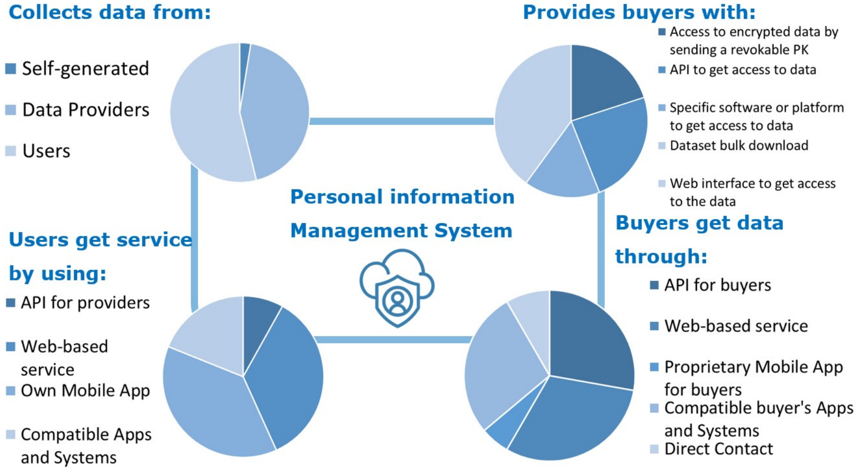


Figure 11: How do PIMS work?

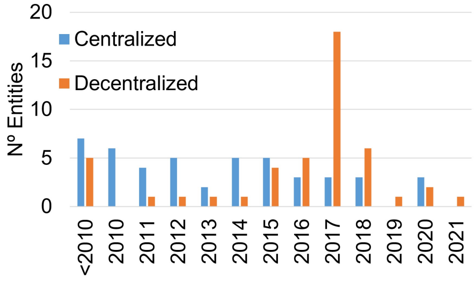


Figure 12: Data management architecture in time

We found that 69% of entities answer this question on their websites, which reflects this is indeed an important issue for them. They claim to be using one or more of the following mechanisms: publishing or sending in advance **free samples** of data to potential buyers; allowing **free access** to part of the data (e.g., some fields of a structured data base); offering a **trial period** in which to have access to a data feed or subscription-based service; providing buyers with a **sandbox** (Battlefin, Otonomo) that lets them experiment with real data before bidding for it or making a purchase decision; offering a live **demo** of their services and their data; or hosting a **reputation mechanism** for buyers to rank data and providers.

4.10 What security measures are taken?

PIMSs and DMs often publish high-level information related to security of data and data exchanges to gain the trust of potential users, be they buyers or sellers. Unsurprisingly, it is *PIMSs* that express the greatest concern about it: most of them include a section dedicated to data security, and address users' and buyers' frequent concerns in this respect. On the contrary, DPs are generally reluctant to give away any information about security, which is considered an internal policy.

Some of the measures taken to secure data include: user authentication and identification; TLS encryption; Anonymization or de-identification of personal information; delivery through revokable DLT decryption keys or public-key cryptography allowing buyers to decrypt an encrypted data stream or dataset; use of temporary URL to deliver data; secure data connectors; tamper-proof data delivery through data signatures and message chaining, which sometimes make use of a blockchain to ensure immutability; or use of specific service and software that certifies the origin of data.

However, DMs still fail to provide a fully effective solution to avoid unauthorized data replication, and protecting data ownership was found to be a key challenge (*Challenge 4*). Moving from providing data to providing services has traditionally been the most commonly accepted recipe to mitigate this risk [65]. For example, ML *niche* DMs look to sell model training services [20], rather than bulk data to train models as *general-purpose DMs* do.

DMSs and PMPs sell data to be used within their systems and services, and heavily restrict outgoing data flows. Extending the scope of controlled environments like *embedded* DMs might be a means to impose severe barriers to data replication and enhance the control of the access to data. Still it needs to be proven whether an open version of such a controlled trust model can be scaled and bootstrapped to the entire Internet as standards like the International Data Spaces [7] and initiatives like Gaia-X [30] are aiming to.

5 RESEARCH ON DATA MARKETPLACES

Once we have presented the results of the survey, this section summarizes the efforts of the research community in defining marketplaces, trading and pricing mechanisms for data.

Other survey papers have been published regarding DMs [55, 64, 66, 67]. Ours is more up-to-date (half of the surveyed entities were founded in 2016 or later), broader in scope, and provides an in-depth analysis of three times more entities than previous works, as shown in Fig. 2. Further - and following our study of 20 of them - this work is also, to the best of our knowledge, the first to address the business models and challenges of personal information management systems (*PIMSs*).

Recent vision papers state the different research challenges envisaged by the research community when building a DM [29]. Not only do they focus on data transactions, but they also emphasize challenges related to discoverability, integration, and transparency, and deal with the systems perspective, too.

An important ongoing research effort is focused on designing marketplaces intended to train ML models. Different value-based data marketplaces have been designed based on this concept, from sellers selecting a price-value from a mix offered by the DM [16], to buyers bidding for data and returning a proportional value in return [1], or collaborative DMs [53]. Some broker prototypes, designed as smart contracts, use cryptography techniques to train - still simple - ML models while protecting data privacy [42]. Other prototypes based on blockchain look to integrate IoT data flows in ML models [75]. In such settings, calculating payoffs according to the accuracy that data brings to a model is a complex problem (*Challenge 5*), often dealt with by approximating Shapley values of data [31, 62]. Some authors have proposed adding a data appraisal layer to select private training data from a ML marketplace [9, 71].

Data pricing has long attracted the attention of the scientific community from different fields [47, 58]. As a result, different schools are applying disparate tools to set arbitrage-free revenue-maximizing prices to data products, often in specific contexts. Such tools include auction design [34, 35], differential privacy [32, 45], pricing queries to a database [15, 41] and quality-based pricing [37, 72, 73]. Finally, other authors focus on personal data within the online ad ecosystem [13, 54, 57], spatio-temporal [4, 5], or IoT sensors data [48].

Despite the ongoing innovation in the market, most theoretical concepts in research papers regarding pricing, privacy-preserving techniques and value-based payment distribution are still under development, and cannot be found in commercial platforms yet. On the contrary, some commercial DMs (IOTA for IoT data, Swash for personal data, or MedicalChain for health-related data) are implementing blockchain-based data exchange platforms, often supported by their own public whitepapers. Some of these start-ups, such as the PIMS Datum, have failed due to the exorbitant costs of storing the transacted data in a blockchain, and most of them restrict the use of DLTs to transactional data.

Next, we summarize the open challenges we highlighted in this paper, and we appoint technologies to address them.

6 OPEN CHALLENGES

Despite its remarkable potential and observed initial growth (50% yearly growth of the number of products offered in AWS and DataRade.ai in 2021 [8]), the market for data is still at its nascent phase. Like all nascent economies, the data economy faces a yet uncertain future. Regardless of which companies and business models finally succeed, we identified some key, intertwined, open challenges related to increasing the *practicality* and *trust* of the ecosystem:

(*Challenge 1*) The *current fragmentation of data markets*, as reflected by the ever-growing large number of companies in market surveys and the variety of data being traded (see Sect. 4.1), makes us think that a consolidation could take place in the future and a new single monopoly or 'niche' data trading champions may arise. Instead of waiting for such champions to solve it for everyone, and given the importance and complexity of the task, solutions must be sought that respect transparent Internet and web governance and expansion principles, including openness, standardization, and layering [44].

(*Challenge 2*) Regarding data economics, there are open problems and unexplored questions related to pricing, as we pointed out in Sects. 4.3, 4.4, and 5. Moreover, a healthy data market requires knowledgeable neutral references (like web services suggesting a range of prices to second-hand car sellers based on the model, year, nº km, etc.) to avoid ending up in a radical and sustained price fluctuation of data products, as recent measurement works show [8].

(*Challenge 3*) Due to the fact that 'data' is an experience good, it is far from obvious for potential buyers to anticipate the value of data in certain settings such as ML tasks [10, 31]. Hence, developing new solutions allowing buyers to first locate [12] and then select better data, which improve the - still primitive - ones presented in Sect. 4.9, is important, too [9].

(*Challenge 4*) Dealing with *ownership* and fighting against piracy and theft of data is of uttermost importance to ensure trustworthiness in data trading. This task is even more arduous when malicious players are able to copy and transmit data at zero cost, and the market lacks a sound notion of authorship, as we pointed out in Sect. 4.10.

(*Challenge 5*) Related to *data provenance*, computing *fair compensations* for providers at scale is an additional challenge for DMs, and especially for PIMSs dealing with individuals (see Sect. 5). In accordance to the research community, such compensations must ideally be based on the value they bring to a specific task or buyer, and DMs must be accountable and transparent about the process [29].

Fortunately, some existing cutting-edge tech will decisively help in overcoming these daunting challenges, namely:

Effective data provenance, even spanning outside of trust ecosystems, may be built upon advances on watermarking [2, 17, 24, 46], hashing [33, 74], trusted execution [63], and network tomography [14].

Usage-based economics may be built upon crowdsourcing [39, 40], cryptography [23], and blockchain-based Non Fungible Tokens (NFTs) [70].

Information Centric Network principles [3] at its data layer provide a base to handle data naming, routing, and in-network storage and replication.

Finally, significant regulatory challenges related to data trading lie ahead, both for competition authorities and *ex ante* regulatory bodies. Due to their market power, tech companies are increasingly under the scrutiny of regulators both in the

US and the EU. Policymakers are currently evaluating the imposition of some degree of data sharing to dominant tech firms in their effort to balance its market power [11]. However, designing such a policy is complex in the case of data assets due to its potential harm to privacy and security.

Within the realm of *personal data*, *protecting privacy* was the main purpose of recent legislation in the EU and the US. New legislative proposals in the EU [27, 68] aim to foster data sharing and '*offer an alternative to data handling practice of major tech platforms*' [28]. Assuming regulatory bodies are able to enforce such regulations, some authors have proposed that individuals are compensated for their personal data [43, 59], while others suggest that entities collecting personal data must be required to act as a fiduciary [22], or even that the mass collection and sharing of sensitive personal data must be banned and prosecuted [69].

7 CONCLUSION

We have catalogued ten different business models in this paper, based on a comprehensive survey that analyzed 104 entities trading data on the Internet. Through this extensive study, it has become clear to us that most of the challenges these entities face have to do with *trust*. On the one hand, sellers express an ambition for absolute control of their data, and demand a strong commitment from marketplaces that data is not replicated and resold, nor used without their authorization. On the other hand, potential buyers need to test data and know its value before closing a transaction, and certify that information comes from trusted data sources.

Not surprisingly, the most successful market players nowadays are horizontally integrated service providers that protect (rather than share) their most valuable data assets. Traditional providers are being challenged by marketplace platforms that work both with data sellers and buyers to facilitate data transactions. It is unclear whether there is a one-fits-all solution, and more recent *niche* coexist with *general-purpose* DMs in the market nowadays. Nonetheless, their business model must still prove feasible, and it is not clear whether specialization is convenient. On the one hand, *niche* DMs have clear advantages over *general-purpose* ones. First, because focusing on certain data space and leveraging their specific expertise let them provide value-added services both to buyers and sellers along with data sharing. Second, because their platform is adapted to the kind of data they trade, and they concentrate their commercial efforts on attracting a specific buyer segment. On the other hand, *niche* DMs target a much smaller market, and the concept of a one-stop-shop for any kind of data is arguably attractive.

Unlike public DMs, *embedded* DMs and PMPs consider data trading more as an add-on to the services they already provide. This *commodification* of data trading has two important competitive advantages. First, they leverage an existing potential customer base on the buy side, which lets them concentrate on finding the right data partners to attract their captive demand. Second, they sell data to be used within a specific environment, which significantly reduces the risk of replication and lets them provide more focused, processed, and thereby more valuable data.

Fighting against the data-for-services dynamics of the Internet is the main challenge of *PIMSs*, provided the rights of new data protection legislation are enforced by competent authorities. They are focusing on gaining the *trust* of users to build a minimum viable base, yet their feasibility is still to be proven. Consequently, they are struggling to make themselves known, leveraging an increasing concern around privacy on the Internet. Conversely, the variety of existing isolated platforms may undermine their trustworthiness. A future consolidation may help them acquire users, though it may well turn the odds against them unless they differentiate themselves from the big '*datalords*'- why trust your data to a *PIMS* instead of Internet service providers? Adopting data *trust* models might be a way to overcome this challenge.

In conclusion, the data economy is a thriving though controversial ecosystem still under development. A huge corporate, entrepreneurial and research effort aims to *de-silo* data and enable a healthy trading of such an important asset, which is key to fully unleash the power of the knowledge economy. In this study we have revealed significant differences between what is working in the market right now and what the market is developing. Through commodifying and specializing data trading, the market is moving away from horizontally integrated monolithic siloed data providers, and towards distributed '*niche*' exchange platforms.

8 ACKNOWLEDGEMENTS

This project has been funded by the European Union's Horizon 2020 Research and Innovation program under the PIMCITY project (Project No. 871370).

REFERENCES

[1] A. Agarwal, M. Dahleh, and T. Sarkar. 2019. A Marketplace for Data: An Algorithmic Solution. In *Proc. of ACM EC'19*.

[2] R. Agrawal and J. Kiernan. 2002. Watermarking Relational Databases. In *Proc. of the VLDB'02*.

[3] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman. 2012. A survey of information-centric networking. *IEEE Comm. Magazine* 50, 7 (2012).

[4] H. Aly, J. Krumm, G. Ranade, and E. Horvitz. 2018. On the value of spatiotemporal information: principles and scenarios. In *Proc. of ACM SIGSPATIAL'18*.

[5] H. Aly, J. Krumm, G. Ranade, and E. Horvitz. 2019. To Buy or Not to Buy: Computing Value of Spatiotemporal Information. *ACM Trans. Spatial Algorithms Syst.* 5, 4, Article 22 (2019).

[6] I. Arrieta-Ibarra, L. Goff, D. Jiménez-Hernández, J. Lanier, and E. G. Weyl. 2018. Should We Treat Data as Labor? Moving beyond "Free". *AEA Papers and Proceedings* 108 (2018).

[7] International Data Spaces Association. 2022. Web page. https://internationaldataspaces.org/. Accessed: Apr'22.

[8] S. Andrés Azcoitia, C. Iordanou, and N. Laoutaris. 2021. What Is the Price of Data? A Measurement Study of Commercial Data Marketplaces. (2021).

[9] S. Andrés Azcoitia and N. Laoutaris. 2020. Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces. arXiv:2012.08874

[10] S. Andrés Azcoitia, M. Paraschiv, and N. Laoutaris. 2022. Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces. *ACM SIGSPATIAL'22* (2022).

[11] C. Biancotti and P. Ciocca. 2019. Opening Internet Monopolies to Competition with Data Sharing Mandates. *Peterson Institute for International Economics. Policy Brief* (2019).

[12] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a Search Engine for Datasets in an Open Web Ecosystem. In *Proc. WWW'19*.

[13] J. P. Carrascal, C. Riederer, V. Erramilli, M. Cherubini, and R. de Oliveira. 2013. Your Browsing Behavior for a Big Mac: Economics of Personal Information Online. In *Proc. WWW'13*.

[14] R. Castro, M. Coates, G. Liang, R. Nowak, and B. Yu. 2004. Network Tomography: Recent Developments. *Stat. Sci.* 19, 3 (2004).

[15] S. Chawla, S. Deep, P. Koutris, and Y. Teng. 2019. Revenue maximization for query pricing. *Proc. VLDB Endow.* 13 (09 2019).

[16] L. Chen, P. Koutris, and A. Kumar. 2019. Towards Model-Based Pricing for Machine Learning in a Data Marketplace. In *Proc. of SIGMOD'19*. ACM.

[17] Z. Chen, Z. Wang, and C. Jia. 2018. Semantic-Integrated Software Watermarking with Tamper-Proofing. *Multimedia Tools Appl.* 77, 9 (2018).

[18] H. Chesbrough and R. Rosenbloom. 2002. The Role of the Business Model in Capturing Value from Innovation: Evidence from Xerox Corporation's Technology Spin-Off Companies. *Industrial and Corporate Change* 11 (2002).

[19] Edward Curry. 2016. *The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches.* Springer International Publishing, 29–37. https://doi.org/10.1007/978-3-319-21569-3_3

[20] M. Dahleh. 2018. Why the Data Marketplaces of the Future Will Sell Insights, Not Data.

[21] DataRade.ai. 2022. Platforms. https://datarade.ai/platforms. Accessed: Apr'22.

[22] S. Delacroix and N. D. Lawrence. 2019. Bottom-up data Trusts: disturbing the 'one size fits all' approach to data governance. *International Data Privacy Law* 9, 4 (2019).

[23] D. Derler and D. Slamanig. 2018. Highly-Efficient Fully-Anonymous Dynamic Group Signatures. In *Proc. of ASIACCS'18*.

[24] G. J. Doërr and J.L. Dugelay. 2003. A guide tour of video watermarking. *Signal Process. Image Commun.* 18 (2003).

[25] EC and IDC. 2021. EU Data Landscape. https://datalandscape.eu/companies. Accessed: Apr'22.

[26] EU. 2016. General Data Protection Regulation (GDPR).

[27] EU. 2020. Data Governance Act (DGA).

[28] EU. 2020. Press release: Commission proposes measures to boost data sharing and support European data spaces.

[29] R. C. Fernandez, P. Subramaniam, and M. J. Franklin. 2020. Data Market Platforms: Trading Data Assets to Solve Data Problems. *Proc. VLDB Endow.* 13, 12 (2020).

[30] Gaia-X. 2022. Web page. https://gaia-x.eu/. Accessed: Apr'22.

[31] A. Ghorbani and J. Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. *Proc. of ICML'19* (2019).

[32] A. Ghosh and A. Roth. 2011. Selling Privacy at Auction. In *Proc. of EC Conference (EC '11)*. ACM.

[33] A. Gionis, P. Indyk, and R. Motwani. 1999. Similarity Search in High Dimensions via Hashing. In *Proc. of VLDB'99*.

[34] A. V. Goldberg and J. D. Hartline. 2003. Competitiveness via Consensus. In *Proc. of ACM-SIAM SODA'03*.

[35] A. V. Goldberg, J. D. Hartline, and A. Wright. 2001. Competitive Auctions and Digital Goods. In *Proc. of ACM-SIAM SODA'01*.

[36] GSMA and AT Kearney. 2018. The Data Value Chain. (2018), 60 pages. https://www.gsma.com/publicpolicy/wp-content/uploads/2018/06/GSMA_Data_Value_Chain_June_2018.pdf Last accessed: Jul 2021.

[37] J. R. Heckman, E. Boehmer, E. H. Peters, M. Davaloo, and N. G Kurup. 2015. A Pricing Model for Data Markets. In *Proc. of iConference 2015*.

[38] N. Henke, J. Bughin, M. Chui, J. Manyika, T. Saleh, and B. Wiseman. 2016. The age of analytics: Competing in a data-driven world. *McKinsey Global Institute* (2016).

[39] C. Iordanou, N. Kourtellis, J. M. Carrascosa, C. Soriente, R. Cuevas, and N. Laoutaris. 2019. Beyond Content Analysis: Detecting Targeted Ads via Distributed Counting. In *Proc. of CoNEXT '19*. ACM.

[40] C. Iordanou, C. Soriente, M. Sirivianos, and N. Laoutaris. 2017. Who is Fiddling with Prices? Building and Deploying a Watchdog Service for E-Commerce. In *Proc. of SIGCOMM'17*. ACM.

[41] P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu. 2012. QueryMarket demonstration: Pricing for online data markets. *Proc. VLDB Endow.* 5 (08 2012), 1962–1965.

[42] V. Koutsos, D. Papadopoulos, D. Chatzopoulos, S. Tarkoma, and P. Hui. 2020. Agora: A Privacy-aware Data Marketplace. In *IEEE Intern. Conf. on Distributed Computing Systems (ICDCS)*.

[43] J. Lanier. 2013. *Who Owns the Future?* Simon & Schuster.

[44] N. Laoutaris and C. Iordanou. 2021. What Do Information Centric Networks, Trusted Execution Environments, and Digital Watermarking Have to Do with Privacy, the Data Economy, and Their Future? *SIGCOMM Computing Comm. Rev.* (March 2021).

[45] C. Li, D. Y. Li, G. Miklau, and D. Suciu. 2015. A Theory of Pricing Private Data. *ACM Trans. Database Syst.* (2015).

[46] X. Liang and S. Xiang. 2020. Robust reversible audio watermarking based on high-order difference statistics. *Signal Processing* 173 (2020).

[47] A. Löser, F. Stahl, A. Muschalle, and G. Vossen. 2012. Pricing Approaches for Data Markets. In *Proc. of the BIRTE*.

[48] W. Mao, Z. Zheng, and F. Wu. 2019. Pricing for Revenue Maximization in IoT Data Markets: An Information Design Perspective. In *IEEE INFOCOM 2019*.

[49] Analysis Mason. 2014. Online data economy value chain. (2014), 75 pages. https://www.ofcom.org.uk/__data/assets/pdf_file/0020/92153/online_customer_data.pdf Last accessed: Jul 2021.

[50] Analysis Mason. 2020. What is the IoT value chain and why is it important? (2020), 6 pages. https://www.analysysmason.com/contentassets/385a5cfa0c1f4aec87dfecc7a19d4e55/analysys_mason_iot_value_chain_oct2020_rdme0.pdf Last accessed: Jul 2021.

[51] D. Moody and P. Walsh. 1999. Measuring the Value Of Information - An Asset Valuation Approach. In *ECIS*.

[52] State of California. 2018. California Consumer Privacy Act (CCPA).

[53] O. Ohrimenko, S. Tople, and S. Tschiatschek. 2019. Collaborative Machine Learning Markets with Data-Replication-Robust Payments. *ArXiv* (2019). arXiv:1911.09052

[54] L. Olejnik, M. Tran, and C. Castelluccia. 2014. Selling Off Privacy at Auction. In *Proc. of NDSS'14*.

[55] W. Org, J. Becker, K. Backhaus, H. Grob, B. Hellingrath, T. Hoeren, S. Klein, H. Kuchen, U. Müller-Funk, U. Thonemann, G. Vossen, F. Stahl, and F. Schomm. 2014. *The Data Marketplace Survey Revisited.* Westf. Wilhelms-Univ., ERCIS.

[56] A. Osterwalder. 2004. The business model ontology. A proposition in a design science approach.

[57] P. Papadopoulos, N. Kourtellis, P. Rodriguez, and N. Laoutaris. 2017. If You Are Not Paying for It, You Are the Product: How Much Do Advertisers Pay to Reach You?. In *Proc. ACM IMC'17*.

[58] J. Pei. 2020. Data Pricing – From Economics to Data Science. In *Proc. of SIGKDD'20*. ACM.

[59] E. Posner and G. Weyl. 2018. *Radical Markets. Uprooting Capitalism and Democracy for a Just Society.* Princeton Univ. Press.

[60] Daria R. 2019. The Future of Data Marketplaces. https://rubygarage.org/blog/big-data-marketplaces. Accessed: Apr'22.

[61] D. Reinsel, J. Gantz, and J. Rydning. 2018. The Digitization of the World - From Edge to Core. *Data Age 2025* (2018).

[62] B. Rozemberczki, L. Watson, P. Bayer, H. Yang, O. Kiss, S. Nilsson, and R. Sarkar. 2022. The Shapley Value in Machine Learning. arXiv:2202.05594

[63] M. Sabt, M. Achemlal, and A. Bouabdallah. 2015. Trusted Execution Environment: What It is, and What It is Not. In *2015 IEEE Trustcom/BigDataSE/ISPA*, Vol. 1.

[64] F. Schomm, F. Stahl, and G. Vossen. 2013. Marketplaces for Data: An Initial Survey. *SIGMOD Record* 42, 1 (May 2013), 12 pages.

[65] C. Shapiro and H. R. Varian. 2000. *Information Rules: A Strategic Guide to the Network Economy.* Harvard Business School Press.

[66] M. Spiekermann. 2019. Data Marketplaces: Trends and Monetisation of Data Goods. *Intereconomics* (2019).

[67] F. Stahl, F. Schomm, L. Vomfell, and G. Vossen. 2017. Marketplaces for Digital Data: Quo Vadis? *Computer and Information Science* 10 (2017).

[68] European Union. 2021. Consultation on the Data Act & amended rules on the legal protection of databases.

[69] C. Veliz. 2021. *Privacy is Power: Why and How You Should Take Back Control of Your Data.* Bantam Press.

[70] Q. Wang, R. Li, Q. Wang, and S. Chen. 2021. Non-Fungible Token (NFT): Overview, Evaluation, Opportunities and Challenges. arXiv:2105.07447

[71] X. Xu, A. Hannun, and L. Van Der Maaten. 2022. Data Appraisal Without Data Sharing *(Proc. of Machine Learning Research)*.

[72] H. Yu and M. Zhang. 2017. Data pricing strategy based on data quality. *Computers and Industrial Engineering* 112 (2017).

[73] D. Zhang, H. Wang, D. Xiaoou, Y. Zhang, J. Li, and H. Gao. 2018. On the Fairness of Quality-based Data Markets. (2018). arXiv:1808.01624

[74] K. Zhao, H. Lu, and J. Mei. 2014. Locality Preserving Hashing. *Proc. of the AAAI Conference* 28, 1 (2014).

[75] K. R. Özyilmaz, M. Doğan, and A. Yurdakul. 2018. IDMoB: IoT Data Marketplace on Blockchain. In *Crypto Valley Conference on Blockchain Technology (CVCBT)*.

A METHODOLOGY

The methodology used to carry out the survey consisted of the following steps:

(1) **Identification of target companies** trading or making business by delivering data. Companies were identified by either searching the web with relevant key words, or by browsing through relevant articles and papers available on the internet.
(2) **Making a quick first assessment** and classifying companies according to the following basic parameters: type of data they are trading, target industry, type of clients, and business model.
(3) **Formulation of a comprehensive set benchmark questions** covering all the aspects we want to answer in this study, and defining a preliminary set of possible answers to each of them. This was further refined during the research process to generate a taxonomy for presenting the results of the benchmark.
(4) **Carrying out a desktop research** to dive deeper into each specific company, answering to the survey questions in a data sheet, and generating a detailed information dossier about the company for consultation purposes in a latter stage as needed.
(5) **Building the data taxonomy by homogenizing the answers to the benchmark questions** for each company and refining the existing taxonomy of answers that allows the comparison of companies.
(6) **Analysis of the results** of this study, both from a technical and a business perspective, identification of key business models and entities operating according to them.

Several iterations were needed in order to come up with a comprehensive set of data trading entities, and fully understand the current market situation.

A.1 Questions

Table 3 summarizes the questions considered in the survey, the different answers we found when studying the different entities, and the section where we present the results.

In addition to answering the former questions, we gathered some general data to classify each entity, understand its maturity, and measure its popularity. These KPIs include the foundation year, country of origin, companies backing the project, the number of employees, how much money they raised, its AlexaRank and its trend in the last months.

A.2 Data collection approach and limitations

Data acquisition was the result of a desktop research based on secondary information available on the internet. As a consequence, the survey relies on information that the target entities are directly publishing on their websites, as well as any related material, such as whitepapers, public videos, product brochures and presentations.

Whenever an answer was not found for any question in the case of a specific entity, "N/A" (meaning *not available*) labels were used. In general, this situation is due to either a lack of information when analyzing such entities, or due to insufficient detail of such information to answer the question. We report the percentage of entities for which we have information in each subsection presenting the results of the survey.

B LIST OF ENTITIES INCLUDED IN THE SURVEY

Table 4 summarizes the list of entities trading data that were analyzed in depth in the survey, including their business model out of the ones defined in Tab. 1. For bigger companies such as SAP or Oracle, the business model reflects the role of their data trading solutions.

In addition to the former companies, the survey included the following entities: AAAChain, Acxiom, Adcolony, Adelphic, Adform, Adition, admaxim, Adobe Advertising Cloud, Adot, AdSquare, adsWizz, adXperience, Algorithmia, Amaxon Mechanical Turk, Amobee, Apervita, ArcGIS Marketplace, Automat, Axonix, Bidtheatre, BigChain, BigToken, Bottos, Bluetalon, CentroBasis, Clearview.ai, Cogito, Complementics, CoverUS, CXSense, Datacoup, Dataguru, DataHub, Datax.io, DataXpand, dbc, Evotegra, Experian, Eyeota, Fyber, Hu-manity, ifeelgoods, IBM, , iExec, Imbrex, ImproveDigital, Informatica Data Exchange, InMobi, LiveIntent, LUCA, Magnite, Mediarithmics, Microbilt, MyHealthMyData, Nielsen, OpenPDS, Opiria Blockchain, Optum Data Exchange, Orderly, OwnData, OwnYourInfo, PickcioChain, PlaceIQ, Pubmatic, Qlik Datamarket, Relevant Audience, Reply.io, Reveal Mobile, ROKU (Oneview), Rubicon project, RythmOne, Smaato, Smartclip, StreetCred, Synchronicity, Tabmo - HAWK, Taboola, TapTap, The DX network, Tremorvideo, Trufactor, Wove, Xandr, xDayta.

After a first quick assessment, we discarded such entities for their subsequent in-depth study and documentation. In particular, we rejected online advertising platforms not offering a private marketplace, concept projects either lacking information or discontinued in time, entities no longer providing service, nor providing any data exchange or data-driven service as such.

Finally, we filtered out some entities those whose business model was already well represented by entities in Tab. 4. For example, we found 2,015 data providers with similar business models, but we only included 35 of them in the survey. We prioritized data or service providers providing clear pricing information. As an exception, we did include every active PIMS we found in the market in order to provide the reader with a thorough overview of this brand-new business model.

Table 4 lists data trading entities analyzed in depth in the survey, and their corresponding business models.

Table 3: Survey questions and taxonomy of the results produced in the survey

Field	Question	Values	Sect.
Type of data	Which kind of data is the entity trading with?	IoT Sensor Data; Personal Data; Geo-located data; Contact data; Marketing; Corporate data; AI / ML models; Human-generated data; Multimedia; Industry; Trading Data; Web data; Automotive-related; Identity data; Healthcare data; Genetic data; Any	4.1
Whose data?	Who are the data subjects?	Individuals; Businesses; Sensors; Any source	4.2
To whom is it sold (B2C, B2B, Any)?	who is provided with data after each transaction? Who is the data consumer?	B2C; B2B; Any	
Targets	Who is the target of the entity? In case of B2B business models, which department or specific industry is the company targeting?	Digital Service Providers; Marketing; Market research; Financial; Automotive; Individuals; Energy, Logistics, Oil & Gas; Healthcare; Retailers; Any	
Pricing mechanisms	Pricing mechanisms available for data being sold by the marketplace	Fixed subscription; Bid by Buyer; Fixed by seller; Auction; Customized; Free; Revenue Sharing; CPM; CPC; %Gross Media spent; Volume-based; Open; N/A	4.3
Actor(s) setting prices of datasets	Who sets the price of traded datasets?	Providers; Platform; Subjects; Buyers; Open	4.4
Payment redistribution mechanisms	How does the platform redistributes payments to data subjects / sellers?	One-to-one; Contribution-reputation-based; N/A	
Data transaction payment	Which payment method and/or currency is used in such transaction?	Fiat currency; Token; Internal credits; N/A	4.5
Platform pricing policy towards data subjects	How are data subjects charged for accessing the platform?	Free; Connection fee; Time subscription; IaaS platform charges; Shipping fees; Freemium; Open; N/A	4.6
Platform pricing policy towards data buyers	How are data buyers charged for accessing the platform?	Free; Connection fee; Subscription; Revenue sharing; IaaS platform charges; Shipping fees; Customized; N/A	
Platform pricing policy towards data sellers	How are data providers / sellers charged for accessing the platform?	Free; Connection fee; Time subscription; Revenue sharing; Freemium; IaaS platform charges; Shipping fees; Partnership; One-off fee; Sales commission; N/A	
Access for providers	How do data providers get access to the platform?	API for data providers; Web-services; Mobile App; compatible DPs' systems; N/A	4.7
Access for buyers	How do data buyers get access to the platform?	API for data buyers, Web-services, Proprietary mobile app, compatible data buyers' systems, direct contact, N/A	
Data sources	Where is data coming from?	internet; Self-generated; Sellers; Data Providers; Users; IoT devices	
Data delivery	How does the DM deliver data?	Access to encrypted data by sending a revokable PK, through an API, through a specific software or platform, by training models with the data; dataset bulk download; Web services	
Data storage	Where is traded data stored?	Centralized public cloud backend, Decentralized private clouds, Centralized private cloud, Data subject's device, Distributed depending on data provider, Centralized backend, DLT, Centralized backend or Public cloud, Decentralized public cloud, Data subject's device and Centralized servers, N/A	4.8
Transaction / Management data storage	Where is the information about transaction stored?	DLT, Public cloud backend, DLT or centralized management, Centralized backend, Distributed depending on data provider, N/A	
Structure of data	Who determines the structure of data to be stored? Can the user share whatever data they want?	Data owner, Application, Data sellers, Data providers and the platform, Platform, Data providers, N/A	
Data preview	How can the data buyer see or test the data before it is transacted?	No way, Free sample Data, Free access, Demo, Trial period, Reputation mechanism, Testing sandbox	4.9
Data Security Measures	How do PIMS/Marketplaces prevent unauthorized access to data while stored? And while it is being moved?	Encryption and SSL, DLT decryption key distribution, Distribution of PK, Special additional measures for PI, Secure storage for PK, User authentication, DLT data replication and immutability protection, Distributed secure data storage	4.10

Table 4: List of entities included in the survey and their business model (links accessed: Apr'22)

Entity	Bss. Model	Entity	Bss. model	Entity	Bss. Model	Entity	Bss. model
1DMC	DM	Data Republic	Emb. DM	HERE	PMP	Qiy Foundation	DME
Advaneo	DM	DataPace	DM	HxGn Content	DP	Quexopa	DP
Airbloc	PIMS+DME	Datarade	DM	Intrinio	DP	Refinitiv	PMP
Aircloak	DME	DataScouts	DP	IOTA	DM+DME	Salesforce	DM
AMO	DM	Datasift (now Fairhair)	SP	Knoema	DM	S&P Global DM	PMP
ArcGIS DM	PMP	Datavant	DME	Kochava	PMP	SAP DM	Emd. DM
Atoka	DP	DataWallet	PIMS+DM	LemoChain	DME	SayMine	PIMS
AWS	DM	Datum	PIMS+DM	LiveRamp	PMP	Skychain	DM
Azure	DM	Dawex	DM	LonGenesis	DM	Snowflake	Emd. DM
BattleFin	DM	Decentr	PIMS+DM	Lotame	PMP	Streamr	DM
Benzinga	DP	DefinedCrowd	DP	Madana	DM	Swash	PIMS
Bloomberg EAP	DP	Demyst	DM	Meeco	PIMS+DME	TelephoneLists	DP
BookYourData	DP	dHealth	DME	MedicalChain	PIMS+DME	Terbine	DM
BronId	SP	Digi.me	PIMS+DME	Mobility DM	DM	The Adex	PMP
BurstIQ	DM	Enigma	DP	MMedia Lists	DP	TheTradeDesk	PMP
Carto	PMP	ErnieApp	Surv. PIMS	mydex	PIMS+DM	Sales Leads	DP
Caruso	DM	Factset	PMP	Narrative	DM	TAUS DM	DM
Citizenme	Surv. PIMS	Factual	SP	Nokia DM	DME	v10 data	DP
Cognite	Emd. DM	Fysical	DP	Ocean Protocol	DME	Veracity	DM
Convex	DM	GeoDB	PIMS+DM	OpenCorporates	DP	Vetri	PIMS+DM
Crunchbase	PMP	Google Cloud	DM	Openprise	PMP	Vinchain	SP
Cybernetica	DME	GXChain	DME	Oracle DMP	Emd. DM	Webhose.io	DP
Datablockchain	DME	Handshakes	DP	OSA Decentr.	SP	Weople	PIMS+DM
Databroker / Settlemint	DM	HAT	PIMS+DME	Otonomo	DM	Wibson	PIMS+DM
Dataeum	PIMS+DM	Health Verity	DM	People.io	Surv. PIMS	Xignite	DP
DIH	DM	HealthWizz	PIMS	Quandl	DM	Zenome	DM