DELIVERABLE D3.1

Design of the user-centric Marketplace

H2020-EU-2.1.1: PIMCity

Project No. 871370

Start date of project: 01-12-2019

Duration: 33 months

Deliverable delivery: 07/12/2020

Deliverable due date: 30/11/2020

Document Information

Document Name: Design of the user-centric Marketplace

WP3 – Understanding the new data economy

Task 3.1

Revision: 01

Revision Date: 26-11-2020

Author: IMDEA and UC3M

Dissemination Level

Project co-funded by the EC within the H2020 Programme		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Approvals

	Name	Entity	Date
Author	Santiago Andrés	IMDEA	26/11/2020
Author	Nikolaos Laoutaris	IMDEA	26-11-2020
WP Leader	Ruben Cuevas	UC3M	26-11-2020
Coordinator	Marco Mellia	POLITO	30/11/2020

Document history

Revision	Date	Modification
Version 1	16 Nov 2020	

List of abbreviations and acronyms

Abbreviation	Meaning
AI	Artificial Intelligence
API	Application Program Interface
CCPA	California Consumer Privacy Act
DM	Data marketplace
DME	Data marketplace enabler
DP	Data provider
DVTMP	Data valuation tool from the market perspective
DVTUP	Data valuation tool from the users' perspective
GDPR	General Data Protection Regulation
IaaS	Infrastructure as a Service
IoT	Internet of Things
ML	Machine learning
PaaS	Platform as as Service
PDK	PIMS development kit
PI	Personal information
PIMS	Personal information management system
UCDE	User centric data economy
UCDM	User centric data marketplace

Executive Summary

The web economy has been revolutionized by an unprecedented possibility of collecting massive amounts of user personal data, which led the web to become the largest data market and created some of the biggest companies in our history. Unfortunately, this change has deep consequences for the privacy of end users, who, deprived of any negotiation power, are compelled to blindly provide their data for free access to services. Personal Information Management Systems - PIMS - aim to give individuals back control over their data, by offering technical means to create transparency in the market.

This deliverable aims to develop further the requirements for a user centric data economy, particularly in the context of a PIMS including data marketplace capabilities. In order to design a marketplace that responds to true market requirements, a survey of more than 75 entities trading with data was conducted in order to understand what the challenges they face are. In addition, research on the state of the art in data marketplace design within the research community is.

As a result, we identified the main data trading challenges, so that our user centric data marketplace (UCDM) design pillars cover key challenges for data marketplaces, particularly:

Challenge		What to do?
1	Data sellers do not know what is a reasonable price for their data	Leverage **market-specific data valuation tools** to provide reference prices for similar information from other marketplaces
2	Data marketplace does not know what is the value of data for the buyer	UCDM to provide different options trading between accuracy and to the buyer tailored to buyer's specific case, thanks to a **data valuation framework**.
3	Data buyers cannot "test" data before they buy it	UCDM **try-before-you-buy** mechanism allows buyers to know in advance the utility of data to their specific task without getting access to them.
4	Data marketplaces are not able to give rewards back to users in accordance to the value they bring to the buyer	DM to support **payoff division functions** that depend on the value that each user brings to a specific model, or just on certain characteristics of traded data (defined by the task or the buyer)

	Challenge	What to do?
5	Data marketplaces are not able to avoid or track data replication by buyers	Include **personalized watermarks or fingerprints** in traded datasets to track data leakages.

Table 1. Data marketplace challenges & design pillars

A high-level design for such pillars is provided, including input, process description and outputs according to the following transaction workflow:

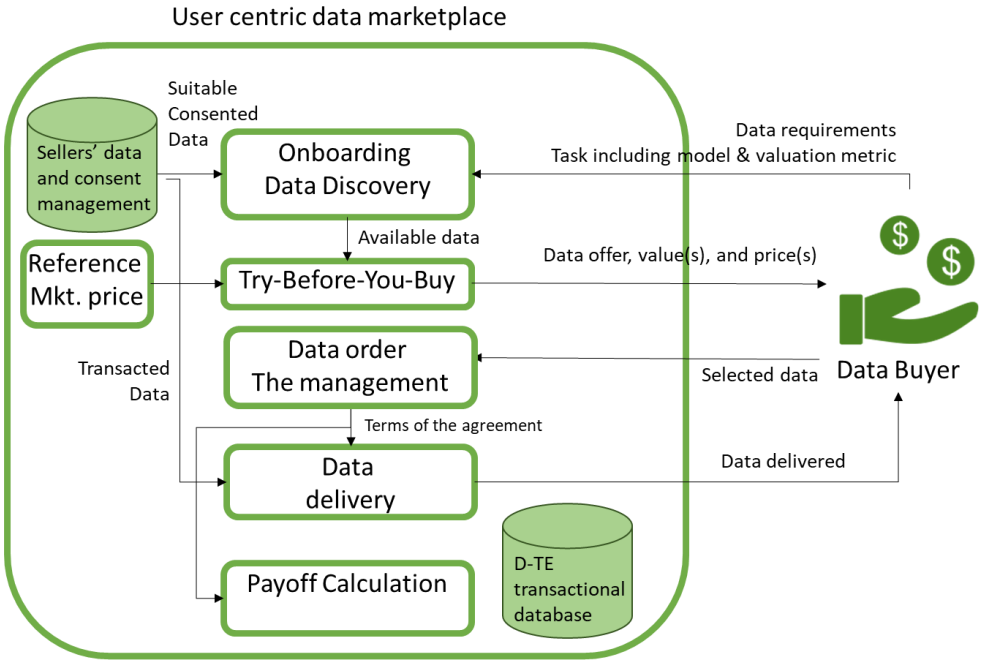

Figure 1. Our proposal of a user-centric data marketplace

Improvements proposed apply to the whole purchasing process, from the moment the data marketplace offers data to a potential buyer for the closure of a data transaction. However, bootstrapping data marketplaces in the data economy is especially challenging, since the market is already dominated by data-driven service providers. We believe a UCDM according to the principles and design stated in this document will have a positive impact on data trading and DM uptake by:

- Trustfully mediating between buyers and sellers, and offering data according to market prices for users to monetize their data,
- Attracting valuable datasets and sellers by discriminating among data depending on the value it has for the specific purpose,
- Incentivizing users to enhance their data to provide more value, hence getting higher compensation,
- Attracting buyers by increasing their trust in the platform as they can know in advance how useful data is for their specific use case, thus allowing them to make wiser decisions and decreasing the risk they are facing in their data sourcing process,

In future deliverables, a design for key UCDM components will be provided, particularly for a reference price tool for the specific case of using personal data in the advertising market, data valuation framework allowing try-before-you-buy and payment division according to different value metrics and a data trading engine governing the whole data transaction workflow for a PIMS data marketplace.

Index

1. Introduction

1.1. PIMCity: Background and objectives

The web economy has been revolutionized by an unprecedented possibility of collecting massive amounts of user personal data, which led the web to become the largest data market and created some of the biggest companies in our history. Unfortunately, this change has deep consequences for users, who, deprived of any negotiation power, are compelled to blindly provide their data for free access to services. Data collection is opaque, fragmented and disharmonic, so that people have lost control over their personal data, and, thus, on their privacy. Personal Information Management Systems - PIMS - aim to give individuals back control over their data, by offering technical means to create transparency in the market. However, so far, they have failed to reach business maturity and a sizeable user base.

In this context, PIMCity aims to offer technical instruments to change this situation by developing an open PIMS development kit (PDK), which will eventually commoditize the creation of PIMS. In addition, a full-fledged PIMS will be developed leveraging the PDK and its benefits shown by populating this EasyPIMS with a significant number of end-users and through collaborations with advertisers and telecom service providers in the market. The final objective of the initiative is to set the basis of a *user centric data economy* (UCDE), meaning that individuals are compensated by companies for their data in proportion to the benefits that such data produce for the overall economy.

1.2. Objectives of D3.1

In deliverable 1.1, a set of requirements for UCDE was preliminarily stated, namely:

- Be cross-compatible among industries, verticals and data types
- Guarantee a fair payoff allocation i) among stakeholders involved in data transactions, and ii) among users contributing to the transacted data
- Allow for flexible data pricing models
- Be scalable in terms of users and transactions
- Provide entities with the right tools to attest data exchanges
- Respect privacy, confidentiality and intellectual property
- Allow for an incremental deployment over the existing technology

First of all, a state-of-the-art analysis will be carried out in order to check to what extent current data marketplaces meet these requirements, and find out what are the challenges for data marketplaces to succeed in the market. For that purpose, a comprehensive benchmark on entities trading with data on the Internet was executed, whose conclusions are summarized in this deliverable. Finally, an intuition about how UCDE contributes to solving those challenges, will be derived from this analysis.

Providing a solution for this whole set of requirements is out of the scope of this project. This deliverable aims to develop further any unaddressed requirement in the context of a PIMS including data marketplace capabilities. Moreover, we emphasize which of them will be tackled in design and implementation activities within PIMCity, and to what extent.

1.3. Structure of the document

In order to meet this objective, this document is structured in the following sections:

- Section 2 summarizes the results of a survey of entities trading with data on the Internet, which was conducted in the scope of the PIMCITY project. It also includes a summary of data marketplace conceptualization efforts carried out by the research community.
- Section 3 states the high-level design and the main features of user-centric data marketplaces, based on the requirements set in D1.1, and taking into consideration the main challenges they face according to our survey.
- Section 4 includes further details and a formalization about how a UCDM works and what the workflow of data transactions is.
- Section 5 provides as a conclusion a summary of how the features defined in our UCDM design will help address the current challenges of data marketplaces, as well as the next steps related to other future deliverables in PIMCITY.

2. Data marketplace survey

"Data", an increasingly essential production factor, just like infrastructure, labor or capital. A myriad of applications in different sectors require huge amounts of information to feed models and algorithms responsible for critical roles in production chains and business processes. Tasks ranging from automating certain functions to facilitating decision-making performed by data-driven organizations are often in need of acquiring such input from third parties. As a result, new entities and novel business models have appeared aiming to match data requirements with the right providers, and facilitate information exchange between organizations.

Before discussing how PIMCITY PDK can contribute to improving the current situation, a desktop research was conducted in order to:

1. What kind of relationships are taking place in the data value network
2. How different entities are selling data
3. What challenges they are facing while doing so

We have also contrasted our findings with PIMCITY partners currently participating in the market, which are also sharing their experience as part of this deliverable. Before presenting the results of the survey, we will make clear what we understand by PIMS, data marketplaces, in contrast to traditional data providers.

2.1. Data Marketplaces and PIMS

Traditional data providers (hereinafter, DP) have either exploited an exclusive access to premium valuable information sources, or either crawled and enriched public information available in the Internet, and leveraged its openness and public data to build successful business models mainly around marketing (Openprise, Lotame, etc), trading (Refinitiv) or business intelligence. DP decouple data sourcing, which they handle directly when needed with their data providers through opaque partnership or specific data acquisition agreements, from the data-driven services they sell to the end-users.

On the contrary, a data marketplace is a two-sided platform that transparently puts in touch sellers with buyers by matching data to solve a buyer's data sourcing needs and, in some cases, facilitates and manages transactions. As demand for data grows, general-purpose marketplaces such as AWS, Advaneo, Data Intelligence Hub or Dawex have entered the scene. They are being challenged by niche marketplaces which cover data sourcing for innovative purposes, such as feeding AI / ML algorithms (Mechanical Turk, DefinedCrowd), IoT real-time sensor data (IOTA, Ocean Protocol / DEX, Datapace), or targeting specific industries and applications (e.g., Caruso for the connected car or Veracity for Energy and transportation industries).

Not surprisingly, some leading data management platforms (Snowflake, Cognite) and niche digital solutions (Carto, Openprise) are integrating data exchange, and even specific data marketplace functionality in their systems. Such built-in

embedded marketplaces provide their users with fit-for-purpose sourcing features that allow them to quickly find and integrate useful data from third parties in their pipeline.

Along with an increasing concern and awareness of Internet users about online privacy, some start-ups have developed solutions to manage and monetize personal data from individuals in the last decade. Such Personal Information Management System (aka PIMS) was recently spurred by new legislative developments such as the General Data Protection Regulation (GDPR) in the EU or the California Consumer Privacy Act (CCPA). Leveraging such legislation, they promise to empower individuals to take control of their personal information (hereinafter, PI) available to Internet service providers and manage their consent to give their data away only to certain entities, or for some specific purposes. Moreover, some PIMS include marketplace functions for users to set a price for their consent, some of which act as trustee of users to monetize their data. Such PIMS would only grant access to PI they control only to those entities who pay a required amount for it and respect the terms and conditions set by the data owner.

For PIMS the data *subject* (the individual who owns personal data the PIMS is trading with) is different from data providers (external entities that hold data from the subject and provide her data to the PIMS). In data marketplaces we generally talk about *sellers* or *providers* (either selling their own data, and/or data enriched from third parties). In both cases we will refer to data *buyers* as the entities interested in acquiring data. Finally, data providers only talk about their providers when they add value to their offering, for example, in the case they signed an exclusive agreement or a partnership with a renowned data provider.

2.2. Data marketplaces survey scope and methodology

2.2.1. Survey questions

The objective of the survey was to find out how different entities are selling data nowadays on the market. For that purpose, we try to find answers to the following questions:

- What is the main business model type the entity is playing in the market?
- Which kind of data is the entity trading with? From whom? Targeting who?
- Where does the information come from?
- How are data subjects charged for accessing the platform?
- How are data buyers charged for accessing the platform?
- How are data providers / sellers charged for accessing the platform?
- Who sets the price of traded datasets?
- What are the pricing mechanisms?
- Does the platform redistribute payments to data subjects / sellers? If so, how?
- Which payment method or currency is used in such transactions?
- How can the data buyer see or test the data before it is transacted?

2.2.2. Methodology

The methodology used to conduct the survey consists of the following steps:

- **Identify target companies** trading or doing business by delivering data, and make a quick assessment. Companies were identified by either searching the web with some key words, or reading articles and papers related to the data economy
- **Make a quick first assessment** and classify companies according to the following basic parameters: type of data they are trading, target industry and type of clients and business model
- **Carry out desktop research** to dive deeper into each specific company, try to answer the survey questions in a datasheet and generate more detailed information dossier about the company for consultation purposes at a later stage
- **Build the data taxonomy by homogenizing their answers to benchmark questions** for each company and refine the existing taxonomy of answers that allows to compare companies
- **Analyze the results of this study**, both from a technical and a business perspective, and prepare the final results to be included in this paper

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

2.2.3. Scope

We checked out more than 75 companies offering data products in order to understand how data is traded these days. No *open data* providers or marketplaces were considered in this selection. Although we agree this is a relevant business model, these platforms offer (often public) data for free, which is not in line with the objectives of PIMCITY and this survey.

Some DP were discarded due to lack of useful information to answer our survey questions. We dived deeper into 48 entities from 16 countries, including 24 data marketplaces, 14 DM enablers and 13 PIMS[1]. The following figure summarizes the final scope of the survey.

[1] Please note that some companies fall in several categories. For instance, some PIMS include data marketplace functionality and are considered in both categories.

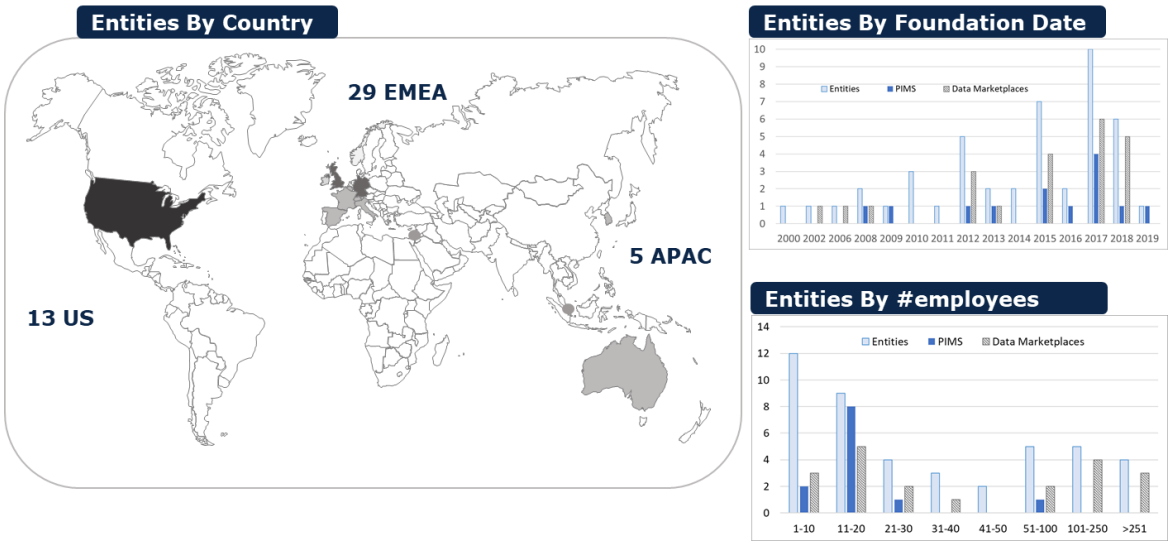

Figure 2. Survey scope summarized

60% of the companies have less than 20 employees and 50% of them were founded in the last five years. In fact, some of them are still start-ups on their way to implement a fully-fledged solution of the products they offer.

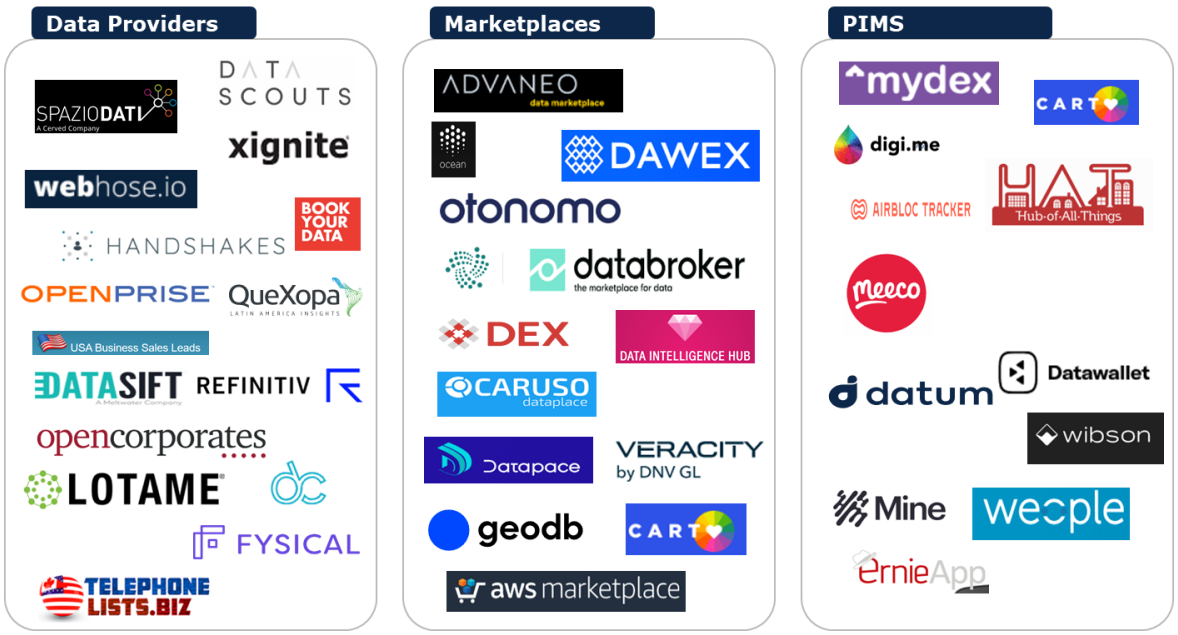

Figure 3. Some companies considered in the survey

2.2.4. Limitations

Data acquisition for this survey paper was gathered by means of desk 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, videos, product brochures and presentations. Whenever an answer was not found to any question in the case of a specific entity, "N/A" (meaning *not available*) labels were used to highlight this situation. In general, this situation is due to either a lack of information when analyzing such entities, or due to insufficient details of such information to answer the question.

2.3. High-level architecture and processes

As part of the survey a series of layers were identified in data marketplace operations according to Figure 4. From the bottom to the top:

1. **The infrastructure layer** aims to provide the basic processing, secure storage and communication functions to the upper layers in the stack.
2. **The platform layer** provides generic APIs and functions to the operational or management processes of the marketplace. There are commercial solutions and PDK in the market that are not intended to directly provide services to the end-users, but to provide a platform with common functions for commercial marketplaces and data providers.

- **The operational layer** provides specific data manipulation and valuation functions that include mainly:
    - data collection or extraction from various sources,
    - data preprocessing, curation, and enrichment process, which may include data fusion with data from other sources to increase its value,
    - data optimization and value calculation for a certain task,
    - operations to perform a secure exchange of data between different entities across the value chain.
- **The management layer** provides sellers' data and transaction management functions. Contrary to the operational layer, it does not manipulate any raw data, but instead:
    - helps data owners catalogue, structure, and price its offer for data,
    - governs any data transaction, including contract management, charging, billing, and accounting processes between entities in the value chain, and
    - supports any audit and tracking of data transactions aimed at increasing the transparency of the data marketplace.

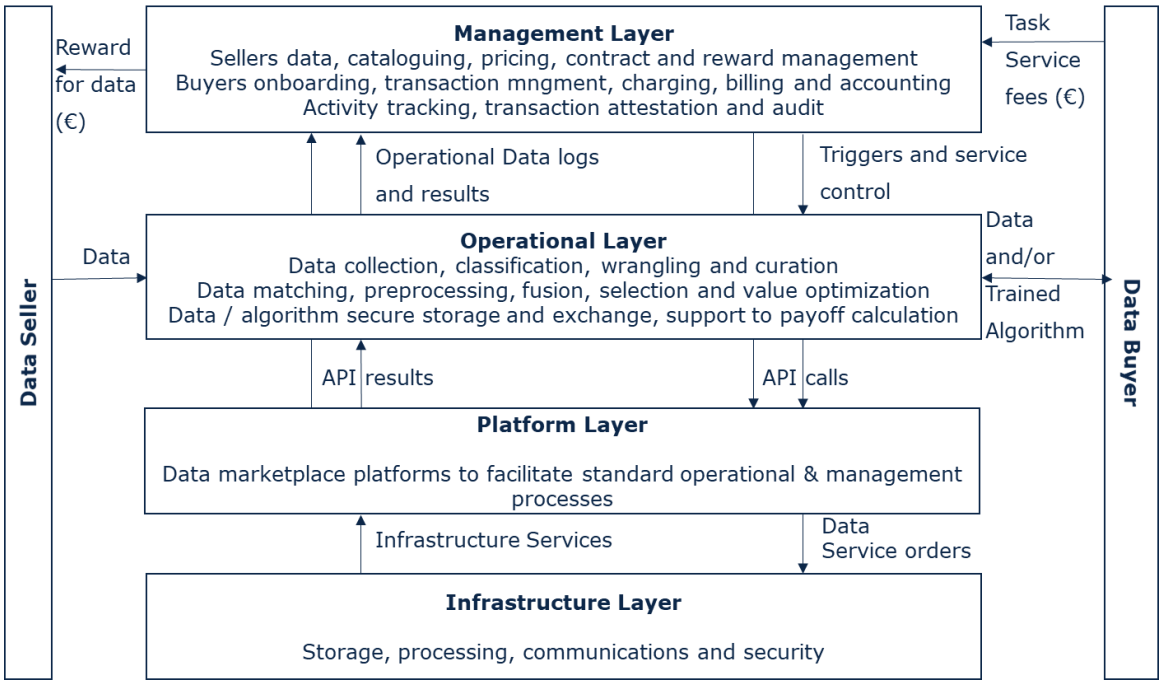

Figure 4. Operational and management layers in entities trading with data

Some PIMS provide marketplace functions in the management layer and all of them allow for cataloguing and controlling personal data from individuals. Other entities in the market are "enablers" meaning that they operate in a wholesale market by providing third parties with functions that facilitate the implementation of operational and/or management processes. Such enablers cannot provide the marketplace of data services to end customers.

Contrary to what one may expect, some marketplaces and PIMS are just putting in touch sellers and buyers, and they let both parties agree on the price out of the platform. They just control that data is not shared with any buyer unless the seller gives her consent. They track data exchanges but they do not influence its economics, such as marketing of a seller's data product, its pricing, billing or charging.

2.4. Summary of survey results

In this section we will summarize the results of the survey for the different questions placed in section 2.2.1. We will group questions of similar nature and we will provide separate results for DP, DM and PIMS to highlight the differences between their business models.

2.4.1. What kind of data is being traded?

The surveyed entities trade all kinds of data, from individuals, companies or both. It is unclear that there is a one-fits-all solution for any data marketplace, since they

trade very different types of data for very different purposes. In fact, niche DM are being successful for certain types of data (e.g. AI/ML, or IoT sensor) and specific industries (e.g. martech, automotive, or energy). Even though most entities are clearly targeting businesses, there is no restriction on individuals to buy data from them.

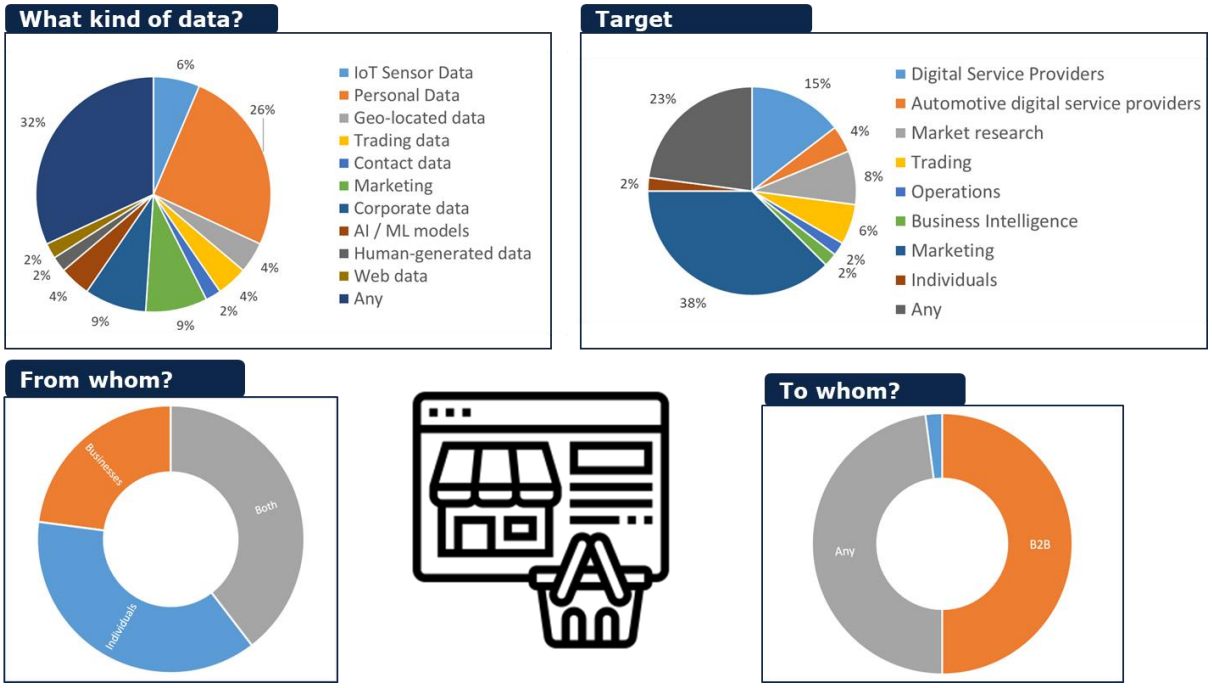

Figure 5. Types of data being traded by marketplaces

2.4.2. How do they charge users?

In general, PIMS (series and info in blue from now on) are free for users, as the main information providers, who are supposed to be attracted by the platform promise of increased privacy and data protection. Buyers are usually free to join the platform, but sometimes they are charged a connection fee or a periodic subscription, as shown in Figure 6.

On the contrary, data marketplaces (hereinafter, series and info in orange) tend to charge both buyers and sellers. A niche DM (Caruso) requires a partnership agreement to be signed by data buyers. The former is subject to similar charging schemes to the ones PIMS use, as shown in Figure 6. The latter are charged for joining or using the platform through one or several of the following schemes:

- Freemium subscription
- Revenue sharing, where the platform keeps a percentage of the total sales

- One-off fees, for instance for connecting to the system

Few niche DM offer partnership models, a model very typical among more traditional data providers, to big data sellers. Finally, PIMS or DM enablers have opted in general for a I/PaaS pricing scheme, and charge entities using the platform for their use of the API or the infrastructure they offer.

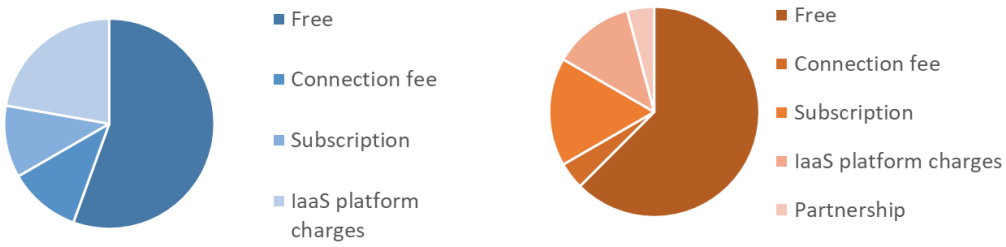

Figure 6. How PIMS and marketplaces do charge buyers for accessing the platform

Regarding sellers, most DM charge sellers a subscription, and/or a percentage of the transactions they manage. Sometimes such subscription is freemium allowing for a growing level of activity or a low percentage per transaction. Few of them offer partnership models to big data sellers, which is usual among traditional data providers. Finally, DME charge buyers I/PaaS-like charges for using their functionality.

2.4.3. How can data be priced?

The following figure summarizes how data is priced and who plays a role at the time of pricing data in each business model. The grey column corresponds to traditional service providers.

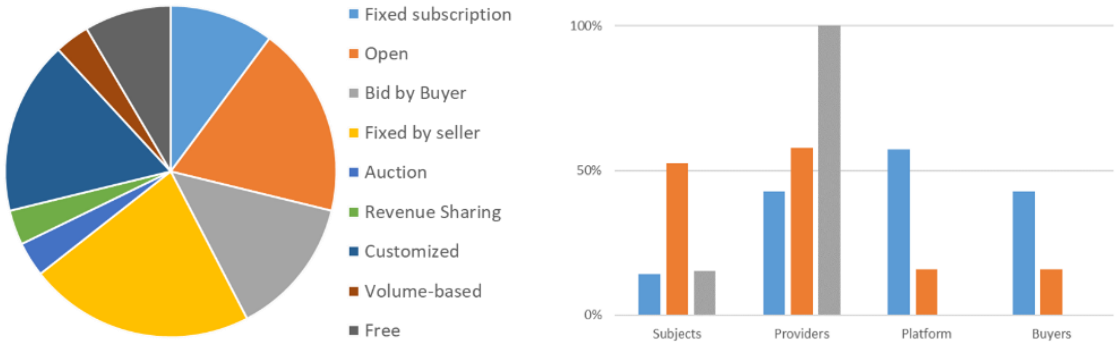

Figure 7. Data pricing mechanisms and players deciding on prices

Regarding how platforms are pricing data that is exchanged in a transaction, as well as which stakeholders are involved in setting such prices. Some of them are flexible in the pricing scheme allowing different mechanisms to be applied as decided by users. Although 23% of entities do not provide clear public information about how data is priced, we consider the following list a comprehensive summary about what mechanisms are used to close transactions:

- **Open**. The entity trading with data does not provide any mechanism to set prices, and it is left for both parties to agree. This approach is usual in entities that enable a secure exchange but do not implement a fully-fledged marketplace.
- **Fixed price**. Buyers pay a fixed price for the data, which could be a lump-sum for a dataset, or a subscription price for accessing a stream or service for a period of time. Most entities providing information about prices (73%) support transactions using fixed prices, and it is by far the most frequent solution adopted by data marketplaces.
- **Bid by Buyer**. Buyers place bids that must be accepted by sellers for the transaction to take place. 30% of entities providing information offer this possibility to users.
- **Volume-based**. Price is fixed depending on the volume of information that is downloaded or accessed. This mechanism is popular among contact data providers, although it is also used by Otonomo to sell data points to automotive app service providers.
- **Customized**: Price is set by the seller case by case depending on who the buyer is and what the data is intended to be used for. In general, the sellers start a transaction by asking such questions to the buyer before price is set. Customized pricing is quite common in data providers and allows price discrimination.
- **Free**. Buyers are able to get data for free, i.e. there is no transaction price. This means that access to data is paid as part of a subscription to the platform (e.g. Carto) or because it is a possibility that the marketplace offers to sellers (e.g. Veracity, DIH).

Although auctions are very popular price setting mechanisms in other fields, it is not common to use it when selling data because of its potential replicability. Only one enabler (Ocean Protocol) names it as a potential mechanism to support when setting prices for data.

Revenue sharing has severe constraints that may discourage its implementation in data transactions. MyDex is the only entity charging transactions using such a scheme: when a buyer purchases the rights to access the personal information of a seller, the platform claims to have rights on the 4% of revenues that such buyer is making out of the individual. Discuss different pricing models and how they affect the interaction between buyers, sellers and the platform.

Regarding who sets the prices of datasets, DP prices are set by the platform. PIMS give more control to users and usually let buyers and users agree on data transaction prices, whereas DMs play a more active role in the process of setting prices for data transactions.

Another interesting topic regarding data transactions is the currency used to pay for them. Whereas data providers traditionally have been billing their services using fiat currency, 50% of surveyed PIMS and 40% of marketplaces decided to use cryptocurrencies to arguably facilitate data transactions. The promised benefits of

this approach would be to increase the speed of transfers, a higher availability if compared to going through banks or establishments, and greater liquidity, fits better the requirements of a real-time data exchange, like the ones trading with IoT sensor data, and actually all startups headed in that direction are betting for cryptocurrencies.

2.4.4. How can data buyers see or test the data before they buy it?

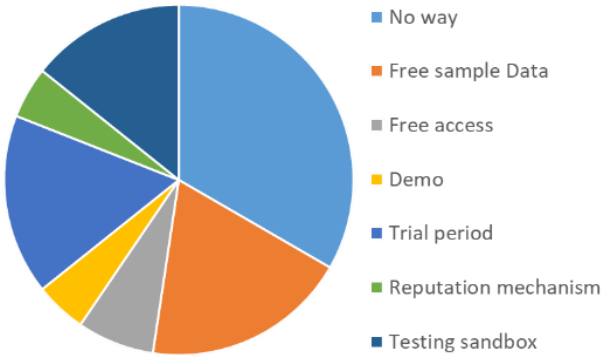

Figure 8. Mechanisms for buyers to test data before they buy

We were not able to retrieve information to answer this question for 25% of the sample. 40% of entities for which we have information do not explicitly offer a way to test data before buying it. The rest of them provide buyers with some sort of visibility over the data they are going to purchase by using one or several of the following mechanisms:

- Publishing or sending in advance a free sample of data (e.g. outdated) for the potential buyer
- Offering a trial period to have access to a data or service
- Offering a demo of the marketplace services or the data, which does not necessarily show the data the buyer is interested in
- Hosting a reputation mechanism by which buyers are able to rank both information and / or data providers
- Providing buyers with a sandbox that lets buyers play with real data before bidding for data or making a purchase decision

Some marketplaces offer buyers the possibility to manipulate data before buying it. This is done through a data manipulation sandbox (Battlefin, Advaneo, Otonomo) which provides buyers with a series of tools to check data, while preventing them from getting a copy of such information.

2.5. Data marketplaces in the research community

Different efforts have been done by the research community in order to define new data marketplaces mainly in two directions:

- IoT-oriented data marketplaces
- ML/AI-oriented data marketplace

2.5.1. IoT-oriented

IoT-oriented data marketplaces (IOTA, Streamr, Datapace) are start-up companies that aim to sell subscription to provider access to IoT data streams, as opposed to bulk datasets sold by traditional marketplaces. Their real-time data transactions are powered by the use of cryptocurrencies and smart contracts to speed-up payments, and most of them leverage blockchain to store management and operational information. They publish their own research concept papers, which give transparency to the way they work on their websites.

2.5.2. ML/AI-oriented data marketplaces

ML/AI oriented data marketplaces are trying to mechanize in a marketplace the process that some niche data service providers are already bringing to the market. (1). is a vision paper that shows how these marketplaces are structured, discusses some design alternatives, and provides useful references on related topics. Most AI/ML-oriented theoretical data marketplace platforms leverage a data valuation framework (2) (3) (4) similar to what we propose to use in PIMCity. In general, they propose that marketplaces train buyers' algorithms or models in a neutral platform by feeding them with their data, and ask for a price for data depending on the accuracy (and, thus the value) the data provides, or even return a trained algorithm yielding an accuracy according to the bid the buyer places in the marketplace, instead of data. The more you pay the higher the accuracy you get in any case.

As data provided to buyers usually benefits from combining different sources, it becomes very relevant the problem of how to fairly split the payment resulting from a transaction among all the sources that contributed to the traded data. Existing marketplaces usually calculate this through simple heuristics, such as the data volume or the number of sources involved. However, simple heuristics are not necessarily tied to the utility of data (5), and consequently, could be considered unfair by sellers. To address this challenge, researchers resort to well-known concepts of game theory to split the revenue among. Most papers propose using the Shapley value (6) for such a task (2), whereas others propose to use *the core* like (7). Finally, some works have studied the dynamics of a data marketplace, and proposed a mechanism that prevents sellers and buyers from postponing their arrival to the marketplace or misreport their costs or values (8).

To the best of our knowledge, there is no practical or commercial implementation of those designs yet, although some digital service providers (cite some of them), provide algorithm optimization and data sourcing services to end customers.

2.6. Conclusions

In conclusion, it is unclear whether there is a one-fits-all solution for data trading, since they trade very different types of data for very different purposes. In fact, niche DM are being more successful by focusing on certain industries (martech, automotive, energy) and types of data (geoDB, Streamr…). Some existing marketplaces are embedded in business optimization tools to enrich enterprise existing data (Qlik, Carto, OpenPrise), or data management tools (Cognite, Snowflake) as part of the solution license. Shall data marketplaces be a standalone player? or is data exchange and trading a functionality to be included in any data-driven service and or data management tool?

In most DM, data pricing is left to sellers and buyers to agree, and in general sellers are responsible to set the price of data to be exchanged. Moreover, taking control of data, and consequently the price for which it is sold is the *leit motiv* of UCDM. Some commercial DM offer professional services to help sellers productize and monetize their data. Due to the nature of data goods, this pricing is increasingly done tailoring the price to the specific client, and the purpose the buyer wants to use data for. In contrast, data marketplaces in the research community define new business models and, for example, try to use bidding as a way to circumvent the problem of tailoring the price.

Existing data marketplaces are decoupled from algorithms. They do not perform any processing on behalf of data buyers unless such service is contracted in addition as an outsourcing or professional service. However, most successful players are horizontally integrated data service providers which exploit free existing data in the Internet or build services on top of an exclusive (and differential) data source. As a result,

On the contrary, new ML/AI data marketplace architectures proposed by the research community try to trade with “accuracy” rather than with data, and make data sourcing somehow opaque to the buyer. They are closer to digital service providers than to traditional data marketplaces, but still it is needed to demonstrate they are able to work in practice.

Are we able to do something in the middle, that complements and helps existing data marketplaces sort out the challenges they face and is implementable in practice?

3. User-centric Data Marketplaces

3.1. Why existing marketplaces do not suffice?

In the light of the conclusions from the survey, there are some relevant challenges that existing data marketplaces face and we will try to address in PIMCITY, namely the following:

Challenge 1. **PIMS users and data sellers do not know what a reasonable price for their data is**. Sellers are usually in charge of setting a price for the data they share. Different studies have been conducted in order to estimate what the value of data is. From the users' perspective, economic experiments, surveys and polls reveal that users really ignore what the value of data is. Only by increasing data trading and exchange will the market build a solid answer to this question.

Challenge 2. **The data marketplace does not know beforehand what the value of data is for the buyer.** Such [2], and companies find it easier to assign value to an improvement in a prediction or the performance of a model rather than directly valuate data used to feed their algorithms. For example, eBay estimated that a 15% improvement on the recommender system translated into 6% increase in revenues[1].

Challenge 3. **Data buyers cannot "test" data before they buy it** and bear the risk of paying for useless data. Although marketplaces provide buyers with a description of data, free samples or let interested parties play with outdated versions of datasets, buyers cannot have access to data before they buy it, or either bid for it.

Challenge 4. **Data marketplaces are not able to give rewards back to users in accordance to the value they bring** to the buyer nowadays. The value of data is inherently combinatorial. It is usually the combination of information which allows buyers to gain rich insights and make the most out of their models. Most data marketplaces sell individual datasets, which are combined and enriched by data buyers. PIMS that offer combined or aggregated data of their users based on an audience specification pay back an equitable reward to users who shared data in a transaction.

3.2. UCDM Concept and Principles

A human-centric data economy requires that individuals are compensated by companies for their data in proportion to the benefits that such data produce for the overall economy. Not only is this a compensation for the negative externality of privacy loss, but a way to improve the quality of data on the Internet, and eventually make the current economic model sustainable, as well. PIMCITY must be designed to follow traditional marketplace approaches, but at the same time to include

[2] https://www.wired.com/insights/2013/04/with-big-data-context-is-a-big-issue/

additional features that enable key functionalities and facilitate UCDM implementation on top of it.

Our proposal for such enabling features is related to the existing approaches in the research community, but differs in scope. Whereas (2) and other authors define a radically new data marketplace concept, which trades with accuracy and mechanizes model optimization services already offered in the market, we intend to extend existing data marketplaces and propose solutions that eventually would help address the former challenges. It is built on top of the following pillars, which are supported by a **data valuation framework** similar to the one proposed by the research community. By using the data valuation framework, an external module is able to retrieve the accuracy of data from the marketplace when used to feed a certain model or algorithm (M). We assume that the buyer knows or is able to calculate the value of data for such a specific task based on such accuracy. Hence, providing information about accuracy is easily translated into value by buyers. As an example, eBay estimated that a 15% improvement on the recommender system translated into 6% increase in revenues (i.e. an increment of $0.54 billion in 2016) (9).

In addition, we leverage the following pillars to design a UCDE:

1. **Reference market prices**. Automated retrieval of prices from the market provides the marketplace with a reference to set prices of data. Such prices depend on the type of data, and the use case. Consequently, the main challenge here is that different tools must be developed for each of them.
2. ***Try-before-you-buy (TBYB)***. We intend to increase the trust of buyers at the time of selecting or buying datasets. By mechanizing the functionality of "*data testing sandboxes*" some commercial data marketplaces already implement, buyers could be provided with an estimation of what the additional accuracy, hence the value, of selecting this or that piece of data will bring to their specific use case.
3. **Value-based payment distribution**. We will allow marketplaces to reward data sellers in proportion to the value their data brings to a user in each transaction, thus incentivizing the provision of high-quality data.
4. **Transparency** is a key principle of a user-centric data marketplace, to sort out the concern of users being treated equally by the marketplace. This is especially difficult when complex calculations are involved in the data transaction process.

3.3. High-Level Architecture

The following picture describes the high-level architecture of what we mean by a user-centric data marketplace (UCDM).

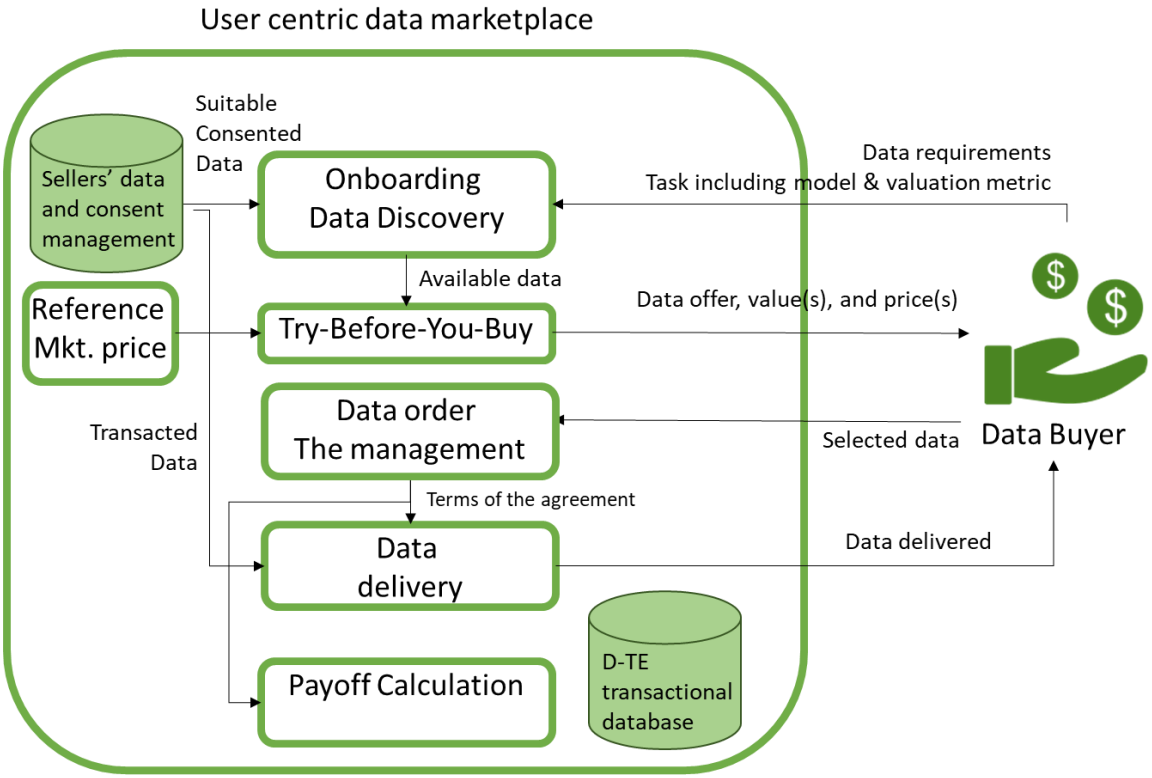

Figure 9. High-level architecture of a user-centric data marketplace

We will assume throughout this deliverable that a buyer who requires the marketplace to provide data for a certain data-driven task shares with the marketplace platform:

- an algorithm or model (M) to be optimized that takes a series of inputs and provides an output y
- Any existing information (D) already under the control of the buyer to be plugged in the model (if not embedded in the model parameters)
- an accuracy valuation function (a) that is able to measure how good or bad the outcome of the model. We will assume it is a similarity function a: y → [0,1]

We will denote by <M, D, a> the task proposed by the buyer.

A UCDM will make use of a parallel neutral environment according to the definitions we give in the **value-based data valuation framework,** using a functionality similar to data sandboxes that existing data marketplaces (e.g., Otonomo, Advaneo or Battlefin) offer. Such a feature will allow the marketplace to execute the buyers' algorithms using sellers' data in order to increase the transparency of the marketplace and attract users (both buyers and sellers).

In the next sections, we introduce the main pillars of UCDM according to the workflow depicted in Figure 1. A deeper discussion regarding each one of them is provided in section 4.

3.3.1. Onboarding. Data discovery

The onboarding process is triggered by a potential buyer's request for data to feed a certain task, and be used for a certain purpose. Such a task may include a request to estimate beforehand the accuracy or value of data from the marketplace in a specific model or algorithm for which the buyer intends to use data. Such a model might be selected from a close catalog provided by the marketplace, or either be provided by the buyer, according to a set of specifications which will be explained on D3.2.

During the onboarding process, the marketplace registers the operation and clarifies any required information before starting any work on the specific operation.

We will assume throughout this deliverable that a data discovery and matching process provides the marketplace with a set S of datasets from sellers in the marketplace suitable for the task that the buyer is sending. Either the data marketplace, buyers, sellers, or a combination of them will be responsible for this process, which is out of the scope of this deliverable. Consequently, we assume that the marketplace is able to identify a set of sources S whose data is suitable to be used in a certain task specified by the buyer, and whose owners have consented to be shared with the buyer and for the specific purpose they are using data for.

3.3.2. Reference market price

One of the main challenges for data sellers is that they face difficulties in setting the price of their personal data. For data sellers, selecting efficient prices requires knowing the level of competition with other data sellers, the willingness to pay of buyers, potential customer lock-ins and other information that affects prices in digital and non-digital markets. Moreover, willingness to pay for data depends on the use case, and on how useful a piece of data is for their specific purpose.

A UCDM must be able to get price references in order to help data owners monetize their data by charging fair prices according to the market value. For that reason, a tool to get reference prices from a market perspective is included in the architecture. However, the main challenge of getting reference prices is that data value has proven to be very different depending on the context and purpose for which it is used.

Within the scope of PIMCITY, such a tool (called data valuation tool from market perspective DVTMP) includes an automated polling of prices in the online advertising industry, which is by far the most frequent use case of personal data nowadays. Such a tool uses historical references from the main audience-based pricing markets, which are those associated with advertising platforms such as Facebook, Instagram, LinkedIN or OpenRTB. The DVTMP module builds specific

crawling tools for a group of advertising platforms to retrieve the value of hundreds to thousands of audiences (i.e., users' profiles based on user's location, demographic information and interests). Audience-based data purchases are one of the most common manners to buy data nowadays. For this specific type of data purchasing process, the DVTMP can provide an estimation of the audience value based on historical pricing associated with that audience on the advertising platforms.

3.3.3. Try-before-you-buy

A relevant problem for data buyers is how to select from suitable datasets for their specific problem. Of course, this problem depends on the kind of information made available to buyers before they make such a decision. UCDM will provide buyers with relevant data so that they make smart purchasing decisions, not only because that would improve the profit of buyers, but also because it will also be positive for the marketplace revenues as it will increase their trustability. Our proposition is based on the former data valuation process, and we call it **try-before-you-buy**. We chose that name since the UCDM will provide buyers with useful information about how data helps them in their specific problem, so that to buy the combination that best suits their needs, just as any of us is able to try clothes in shops. This allows a handful of purchasing and bargaining schemes as we state later in this deliverable.

UCDM may choose to disclose to buyers a combination of information about how suitable data is for their specific case before they make a purchase decision:

- Available datasets suitable to buy
- Accuracy a(s) achieved by individual datasets or combinations of them
- Prices of datasets or combinations of them

3.3.4. Data order management

With the information shared by the marketplace, the data buyer may select one or some datasets out of the offer, and issue a data order for the marketplace to manage. A series of checks are triggered, e.g., ensuring that datasets selected were among the ones offered by the marketplace, meaning that their owners have consented to their information being shared.

It also triggers:

1. The billing, charging and payment process for that order
2. The delivery of data to the buyer
3. The calculation of the compensation to sellers that contributed to the transaction

3.3.5. Value-based payment distribution

As data transactions in UCDM may likely involve several datasets, the problem of how to split any subsequent payment between the sources that somehow

contributed to the transacted data becomes very relevant. We call this the **payment distribution function**, and it will be used by UCDM at least,

1. when agreeing on a transaction, to provide each seller with an expected reward before accepting a transaction that involves a combination of data from different sellers, and
2. once a transaction is closed, to distribute the resulting payment among the sellers that contributed to the traded dataset.

A UCDM must provide a handful of tools for that, including a solution that allows marketplaces to split data depending on how useful each source of data was on that specific problem. We call this value-based payment distribution.

3.3.6. Data delivery

Once a transaction is closed, the set of purchased data must be delivered to the buyer. We assume that as a result of a transaction, the buyer gets a bulk dataset. Some data marketplaces claim that no data should be exchanged, but the buyer should be rather provided a trained model to avoid data leakage.

In our architecture, we propose tracking data leakage by placing a fingerprint on every dataset delivered by the UCDM. The purpose of such a security measure is the following: Were any confidential piece of data belonging to sellers or users of the marketplace be found, the watermarking module would be able to find out who was the recipient of that version of data by reading the watermark inserted on it.

3.3.7. Transparency and attestability

Finally, all the former processes and functions must allow users, be they buyers or sellers, to track the activity of the marketplace. UCDM must be transparent and therefor create logs that enable their users to attest every transaction they are involved in. Such transparency must also be given while preserving the privacy of other users of the marketplace.

3.4. Relationship with other PIMCITY components

Date marketplaces are not necessarily tied to PIMS. In fact, some of them trade with non-personal data. The former UCDM architecture and processes were intended to be used in a generic data marketplace architecture, although they are applicable in the case of a PIMS.

An additional challenge that PIMS face is that the scalability of the number of sources: whereas usually general-purpose marketplaces trade with aggregated datasets including different data, PIMS monetize personal data associated to individuals. Hence, the number of data sources is expected to be higher if the PIMS trades with datasets aggregating data from different people.

Moreover, sellers usually join data marketplaces because they want to monetize their data and so, offer it to third parties. This is not necessarily true in the case of a PIMS. Most users might actually join a PIMS because they want to take control of their data and select who has access to it. Consequently, consent management to decide which data could be used for what is especially relevant in the case of a PIMS. Consequently, data discovery must be executed in two steps:

1. Find out which users have provided data that suit the needs of the buyer (e.g. location data to know how many people are close to a specific location at a certain time for marketing purposes).
2. Filter out those users who did not consent to their data to be used for such a specific purpose, or to be sold to a specific buyer.

We will call *target audience* to a set of individuals whose data is eligible to be used in a certain transaction.

The following interfaces have been identified with the different modules of a UCDM if it needs to be implemented on top of a PIMS:

- All data to be retrieved from the personal data storage module, which is the repository where all the information is stored in the PIMS. Such a module will also ensure that any personal information provided to other modules has been consented by its owner to be used for such a specific purpose and for the specific buyer or type of buyer.
- An optional privacy-preserving querying module could be considered in order to build aggregate datasets for such tasks where there is no need to access personal information of the target audience, but aggregated data from them.
- Data reference prices from the market will be useful by UCDM to facilitate price setting and manage sellers' expectations.
- The UCDM must implement a value-based data valuation framework, to orchestrate the whole trading and data valuation process. This will play an important role in supporting the try-before-you-buy and in splitting revenues among the sellers that contributed to the various sources.
- A UCDM will place a watermark and/or fingerprint in all data to be sent to buyers once a transaction is closed. Consequently, an interface with the data provenance module must be provided in order to fulfil this requirement.

4. User-centric data marketplace components

In the next sections, we define and describe the different functions of the UCDM, including their inputs, process description, place in the workflow and outputs.

4.1. Definitions

The following picture describes how the UCDM can use a data valuation framework in order to provide transparency in data transactions:

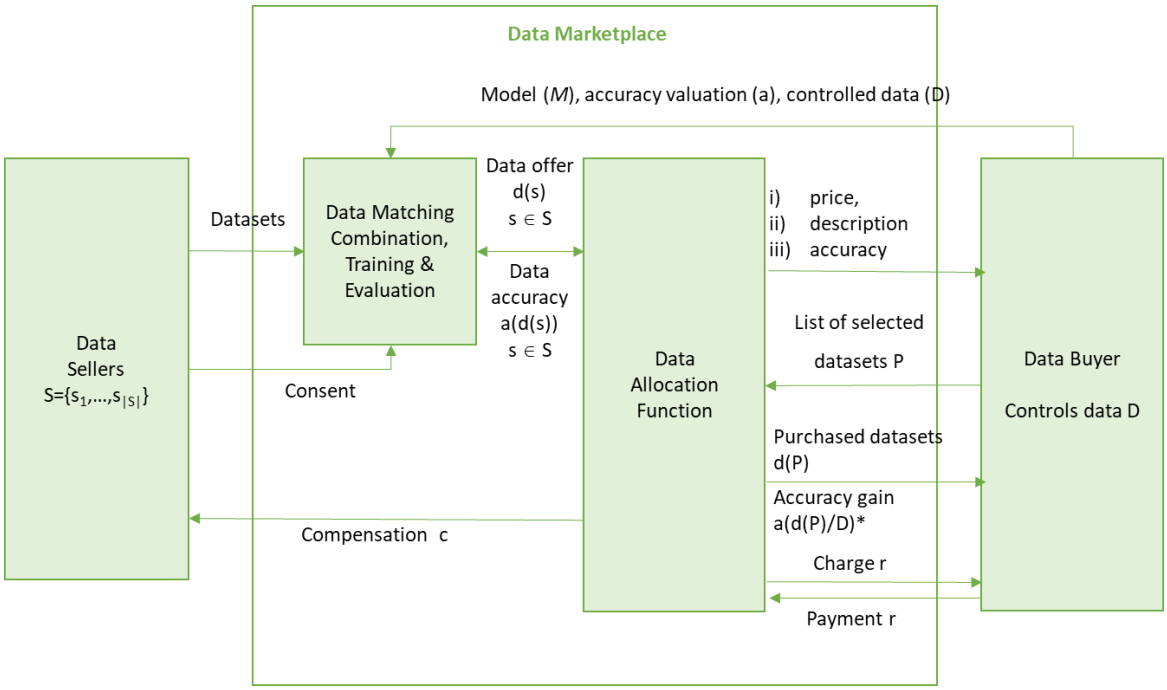

Figure 10. Data purchasing processes based on data valuation framework

As the value of data is closely related to the task context, we will suppose a buyer brings a data-driven task, which is defined by

i) a model M (e.g. a ML model to classify websites by its content),
ii) an accuracy function a, such that a(d) $\in$ [0,1] denotes the accuracy achieved by the model M after feeding it with dataset d, and
iii) any pre-existing dataset D the data buyer has used to train or feed the model with.

We also assume that the marketplace is able to find a series of suitable datasets to feed such a model. S denotes the set of sellers, d(s), s $\in$ S stands for the dataset offered by a seller s, and d(S') the combined dataset of all sellers s $\in$ S', i.e.

$$d(S') = \bigcup_{s \in S'} d(s), S' \subseteq S$$

We will suppose the marketplace knows and informs buyers about the prices p(s) of eligible suitable datasets and, is able to calculate the accuracy a(S') for any subset of all potential combinations of sellers in S, S' $\subseteq 2^S$.

The marketplace is able to inform the buyer not only about eligible datasets or prices, but about the accuracy that individual datasets, or combinations of datasets are able to bring to {M, a, D}. With that information, the buyer is able to know the value that datasets will bring to the model, and identify a suitable set of sellers or datasets to buy from the marketplace. Let us define as P $\subseteq$ S as the set of eligible datasets or sellers selected by the buyer, which yields an accuracy a(d(P)/D).

A buyer is allowed to:

1. Either directly purchase individual datasets at the asking price. In that case, the marketplace generates |P| one-to-one transactions for every element in P.
2. Or place a bid b for P below the asking price (sum of prices of all datasets in s). In that case, it will use the payment distribution function to notify each seller involved in the operation, and get their acceptance. If all sellers involved accept, then the transaction is closed for b.

As a result of a transaction,

i) the marketplace delivers the corresponding data d(P) to the buyer, and charges an amount r which is b, or:

$$r = \sum_{s \in P} p(s)$$

ii) Compensation c(s) is paid to the sources that contributed to the transaction, which in the case of an accepted bid for d(P) will require to fairly split part of the revenues among the sellers that contributed to that transaction. We assume that the marketplace will keep part of the payment in exchange for the service provided to both parties.

4.2. Data valuation framework

The following figure and table summarize how the data valuation framework works: Its design was intentionally agnostic from the underlying model or algorithm and

allows to get the accuracy of the model when fed by any subset of eligible data provided by the marketplace.

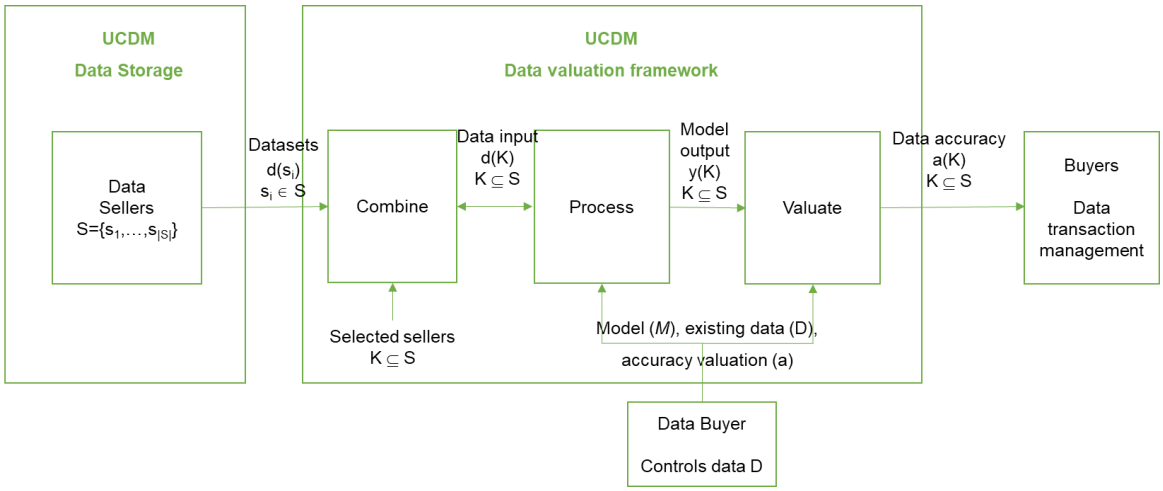

Figure 11. Value-based data valuation framework

Inputs	1. Transaction ID 2. <M, D, a> the task proposed by the buyer 3. Set of suitable sources S = {s_1, … , $s_{\|S\|}$} that have relevant data for the task, as a result of the data discovery process
Description	The data valuation framework is able to execute a model or algorithm provided by a buyer by feeding it with data from sellers in the marketplace in a neutral platform. This is done both preserving the intellectual property of the buyer who shares code, and protecting data sellers from their information being exchanged unless the buyer pays for it.
Output	a(S') for any subset of all potential combinations of sellers in S, S' $\subseteq 2^S$
Workflow	The data valuation framework is an underlying block in the data marketplace transactional workflow. It is used by: 1. Try before you bid process to get information of the accuracy achieved by individual datasets 2. Payment division function to break down a payment r in value-based compensations to users whose data was used in the transaction 3. Price update function to discriminate on the value of data

Table 2. Data valuation framework summary

This framework is intended to be implemented on top of a neutral data sandbox like the ones already in place in some data marketplaces such as Advaneo, Battlefin or Otonomo. Such a sandbox must provide a protection for intellectual property shared by the buyer, and make sure there is no data leakage to the buyer, who will have

only access to the final purchased data. Designing and showing how this value-based data valuation framework works is part of D3.2.

4.3. Try-before-you-buy purchasing process

Try-before-you-buy takes advantage of the data valuation framework and allows the marketplace to disclose information about the accuracy of datasets to buyers on their specific task <M, D, a> **before they make a purchase decision**. Depending on the purchasing mechanism to be used, the marketplace may decide to present buyers:

- Either the prices of individual datasets, combinations of them or no price at all
- Accuracy of individual datasets, combinations of them or no accuracy at all

A marketplace may decide to set individual dataset prices, which is the current situation in marketplaces according to the survey, or prices on combinations of datasets according to the accuracy they provide. (3) and (2) set prices not to datasets but to accuracy, and are able to provide different “versions” of a combination of datasets by adding controlled noise. That way, marketplaces are able to generate a version of data delivering a specific accuracy according to the payment made by the buyer. Such accuracy will be bound by the maximum she is able to reach by using all existing data.

If no price is shown to buyers, they are allowed to bid to get suitable data to their task, according to the value such data has for them. Therefore, a marketplace could implement a bidding process that delivers data according to the bid and the prices of data (2), or accept bids for individual datasets and decide case by case.

Using this functionality, the marketplace will be able to implement different alternatives depending on the amount of information disclosed about prices and accuracy:

<table>
<tr><th>Accuracy
Price</th><th>No accuracy info</th><th>Only for individual datasets</th><th>For individual and combinations of datasets</th><th>Full range of accuracy options (pricing accuracy)</th></tr>
<tr><td>No price</td><td>Blind bid</td><td colspan="3">Bid for accuracy
(2)</td></tr>
<tr><td>Only individual datasets</td><td>Current situation
Stepwise Purchase & try</td><td rowspan="2">Try-before-you-buy (PIMCITY)</td><td>Full information</td><td rowspan="2">Accuracy purchase</td></tr>
<tr><td>Individual and / or combined datasets</td><td>One-shot purchase & try</td><td>Full information with combined prices</td></tr>
</table>

Table 3. Different alternatives in the value-based purchasing process

We will assume in this deliverable that prices are disclosed to buyers, together with a description of datasets, its metadata and perhaps some volume metrics, such as the number of sources or data rows included in the dataset. This is the most common situation in marketplaces nowadays. However, price is not enough to make good data purchasing decisions, as it might turn out that the most promising datasets are not the most useful ones for the buyer's specific use case (5).

We will use try-before-you-buy to improve the information that is available for buyers to decide on which datasets to buy. In general, there are two stages in such a purchasing process:

1. **Data offer**: The marketplace discloses price and/or accuracy information about eligible datasets, or combinations of them, for buyers to make purchase decisions. This falls within the data marketplace side.
2. **Purchasing process**: The buyer selects which alternative(s) presented suit her needs, and asks for it / them. This process is run by the buyer and will select a subset of datasets from the data offer the buyer is interested in. The buyer will then issue data orders to get access to such data.

4.3.1. Data offer (marketplace side)

The following table formalizes the way data offer processes work in a UCDM:

Inputs	1. Transaction ID 2. <M, D, a> the task proposed by the buyer

	3. Set of suitable sources S = {s_1, … , $s_{\|S\|}$} that have relevant data for the task, as a result of the data discovery process 4. Dataset prices p(s) 5. Accuracy a(s) achieved by datasets s $\in$ S, or combinations a(S') of datasets S' $\subseteq$ S in the task <M, D, a>
Description	Within the purchasing process, the marketplace provides buyers with a set of options of a "menu" which the buyer can select in the purchasing process. It states for every option the data d_i , a description of data D_i, its price p_i, and the accuracy a_i it brings to the task <M, D, a>.
Output	A set of tuples O = {<d_i, D_i, p_i, a_i>}
Workflow	The marketplace generates the data showroom as soon as the discovery process ends, and sends such information for the buyer to make purchase decisions. It then triggers the purchasing algorithm of buyers to select the best option from the menu given by the marketplace.

Table 4. Data offer summary

The implementation of the data valuation framework allows marketplaces to implement different data offer alternatives according to table 1. Designing and showing how this value-based data valuation framework works is part of D3.2.

4.3.2. Purchasing process (buyer side)

Buyers will start their purchasing process once they are given a set of options by the marketplace. Buyers will follow a different strategy depending on the amount of information disclosed by the marketplace, always looking forward to maximizing their profit. For example, they will make different purchase decisions if the marketplace only shows prices for datasets and their volume, or if they are also provided the individual accuracy they bring to her model.

Inputs	1. Transaction ID 2. Set of options provided by the marketplace: $O = \{<d_i, D_i, p_i, a_i>\}$ 3. Accuracy target for the buyer 4. Value for accuracy function v(a) 5. Risk appetite for the buyer
Description	Within the purchasing process, the try-before-you-buy process describes how buyers choose among eligible datasets looking forward to maximizing their profit, defined as: $\pi = v(a) - p(P)$ where v(a) is the value achieved after the purchase of a set of options $P \subseteq O$, and p(P) is the sum of prices of options in P. In general, the buyer will lower the risk of the purchase operation if it is executed in rounds, selecting oe option at a time.
Output	Set $P \subseteq O$ of selected purchase options
Workflow	The purchasing process using try-before-you-buy starts after the data discovery process and guides buyers into selecting the right datasets for their needs, and the marketplace provides the set of options to buyers. It triggers both data delivery and payment distribution processes. It stores all the details of the transaction in the transparency log and database.

Table 5. try-before-you-buy – purchasing algorithms summary

Not only will try-before-you-buy benefit the buyers by allowing them to select the best datasets, but contribute to increasing their trust in the marketplace. As a result, we also expect this is beneficial for the marketplace, which will increase sales, and consequently for the sellers, who will receive higher rewards, as well.

4.4. Payment distribution modules

Once a transaction is closed, it may likely involve several pieces of data combined into a single dataset. As a result, the UCDM must find a fair way to split any revenue generated among the sellers that contributed their data to the transaction.

Inputs	1. Transaction ID 2. <M, D, a> the task proposed by the buyer 3. Set of suitable sources $S = \{s_1, \dots, s_{\|S\|}\}$ that have relevant data for the task, as a result of the data discovery process 4. Set $P \subseteq S$ of datasets selected to be purchased 5. Accuracy a(s) achieved by datasets $s \in S$, or combinations a(S') of datasets $S' \subseteq S$ in the task <M, D, a>
Description	Within the purchasing process, the try-before-you-buy process describes how buyers choose among eligible datasets looking forward to maximizing their profit.
Output	Compensation c(s) as a percentage of the reward resulting from the transaction to be paid to each individual contributing to it.
Workflow	The purchasing process using try-before-you-buy starts after the data discovery process and guides buyers into selecting the right datasets for their needs. It triggers both data delivery and payment distribution processes. It stores all the details of the transaction in the transparency log and database.

Table 6. Payment distribution module summary

Designing this module is part of D3.2, which will be demonstrated as part of WP3.

4.5. Data delivery

Once a certain dataset (or a combined dataset) is selected by the buyer and the data transaction is closed, it is delivered to the customer. We assume the marketplace will provide the buyer with a bulk download of the selected dataset. Any dataset must be inserted a *fingerprint* before being sent to the buyer. Such a process must allow the marketplace to find out who leaked any authorized piece of information by accessing the watermark key and checking which transaction generated it in the UCDM transaction database.

Inputs	1. Transaction ID 2. Set P of purchased datasets
Description	After a transaction is closed, the purchased datasets must be delivered to the buyer. We assume that as a result of a transaction, the buyer gets a bulk dataset.
Output	Dataset d(P) delivered to the buyer, including a *fingerprint* capable of identifying the buyer in the case of a data leakage.
Workflow	The data delivery process is triggered by the try-before-you-buy process. It triggers a watermarking process in the data provenance module, and stores the required watermarking data in the transaction log and transparency database.

Table 7. Data delivery summary

4.6. Transaction logs & management

One of the pillars of UCDM is transparency. Their users, be they sellers or buyers, must be able to get all sorts of information about the transactions in which they were involved. For that reason, it is key that the UCDM keep track of every step of the operation, and that every function stated before stores in an indexed transaction log every input, intermediate result and output of every transaction. This includes the following information:

- An identification of the transaction, which must be coherent throughout the process of such transaction
- An identification of the task proposed by buyer
- Identification of the eligible set of sources $S = \{s_1, \dots, s_{|S|}\}$
- Set of options provided by the marketplace: $O = \{<d_i, p_i, a_i>\}$
- Set $P \subseteq O$ of options selected by the buyer
- Accuracy a(s) achieved by datasets $s \in S$, or combinations a(S') of datasets $S' \subseteq S$ in the task <M, D, a> used to calculate payment division function
- Permutations of sources used to calculate such payment division
- Compensation c(s) as a percentage of the reward resulting from the transaction to be paid to each individual contributing to it.

Access to the logs and management database must be done in a privacy-preserving way. For example, at the time of a seller getting access to the relative valuation of their data, no information that allows to identify other sellers must be shared. For example, a seller may have access to the information needed in order to compare the compensation received as a result of a transaction with the "average" seller, so that he is able to know why such compensation was higher or lower than the average, and ideally why. But such a seller will not have access to any comparison with any other specific seller that could lead to disclosing relevant personal information or even lead to identifying the other seller.

5. CONCLUSION AND NEXT STEPS

PIMCity UCDE concept helps address key existing challenges for data marketplaces nowadays. The following table summarizes the main data marketplace challenges and maps to solutions proposed within the UCDM architecture in order to help in solving each one of them:

	Challenge	What to do?
1	Data sellers do not know what a reasonable price for their data is	DM to help with this challenge using the DVTMP that provides reference prices for similar information from other marketplaces. Such references depend on the type of data and the use case. Within the scope of the project, a tool is developed for advertising, which is the most extended use case for personal data.
2	Data marketplace does not know what the value of data for the buyer is	UCDM purchasing mechanism allows buyers to pay for data according to the value they provide to their specific task, and provides buyers with different options in terms of price vs utility / accuracy of purchased data.
3	Data buyers cannot “test” data before they buy it	DM to provide buyers with relevant privacy-preserving insights about the utility of data for their specific purpose before they make the purchase decision through try-before-you-buy processes. DM to provide different options trading between accuracy and to the buyer tailored to buyer’s specific case.
4	Data marketplaces are not able to give rewards back to users in accordance with the value they bring to buyers	DM to support payoff division functions that depend on the value that each user brings to a specific model, or just on certain characteristics of traded data (defined by the task or the buyer). DM to provide different options trading accuracy and processing requirements.

	Challenge	What to do?
5	Data marketplaces are not able to avoid or track data replication by buyers	Include personalized watermarks in traded datasets to track data leakage. Avoid trading raw datasets but aggregated and anonymized data, or trained models instead. Or either sell access to data using a temporary PK so that DM can control who asks for access and how many times. That would require a change in the Internet paradigm.

Table 8. Summary of data marketplace challenges, and solutions proposed by PIMCITY

Improvements apply to the whole purchasing process, from the moment the data marketplace offers data to a potential buyer for the closure of a data transaction. Bootstraping data marketplaces in the Data Economy is especially challenging since the market is already dominated by data-driven service providers. As a complex two-sided platform, it must attract both buyers and sellers, to enter a virtuous growth circle. According to our survey, DM are spending effort and money to attract a sufficient set of data sellers, and then try to convince as many buyers as possible to start purchasing these datasets.

We believe a UCDM according to the principles and design stated in this document will have a positive impact on data trading and DM uptake by:

- Trustfully mediating between buyers and sellers, and offering data according to market prices for users to monetize their data,
- Attracting valuable datasets and sellers by discriminating among data depending on the value it has for the specific purpose,
- Incentivizing users to enhance their data to provide more value, hence getting higher compensation,
- Attracting buyers by increasing their trust in the platform as they are able to know in advance how useful data is for their specific use case, thus allowing them to make wiser decisions and decreasing the risk they are facing in their data sourcing process,

In future deliverables, a design for key UCDM components will be provided, particularly for:

- A reference price tool for the specific case of using personal data in the advertising market,
- Data valuation framework allowing try-before-you-buy and payment division according to different value metrics
- A data trading engine governing the whole data transaction workflow for a PIMS data marketplace

6. References

1. *Data Shapley: Equitable Valuation of Data for Machine Learning.* **Zou, Amirata Ghorbani and James Y.** Long Beach, California : s.n., 2019. International Conference on Machine Learning.

2. *A Marketplace for Data: An Algorithmic Solution.* **Anish Agarwal, Munther Dahleh, Tuhin Sarkar.** s.l. : ACM, 2019. Conference on Economics and Computation. pp. 701-726.

3. *Towards Model-based Pricing for Machine Learning in a Data Marketplace.* **Lingjiao Chen, Paraschos Koutris, Arun Kumar.** s.l. : ACM, 2019. International Conference on Management of Data. pp. 1535-1552.

4. *Data Market Platforms: Trading Data Assets to Solve Data Problems.* **Raul Castro Fernandez, Pranav Subramaniam, Michael J. Frankin.** 2020. VLDB Endowment Vol. 13 No. 11.

5. *Computing the Relative Value of Spatio-Temporal Data in Wholesale and Retail Data Marketplaces.* **Santiago Andrés, Marius Paraschiv, Nikolaos Laoutaris.** 2020.

6. *A Value for n-Person Games.* **Shapley, Lloyd S.** s.l. : RAND Corporation, 1952.

7. *If You Like Shapley Then You'll Love the Core.* **Tom Yan, Ariel D. Procaccia.** 2020.

8. *Data Markets wit Dynamic Arrival of Buyers and Sellers.* **Moor, Dmitry.** Phoenix : ACM, 2019. ACM Economics of Networks, Systems and Computation. p. 9.

9. *Optimizing similar item recommendations in a semi-structured marketplace to maximize conversion.* **Y. M. Brovman, M. Jacob, N, Srinivasan, S. Neola, D. Galron, R. Snyder and P. Wang.** s.l. : ACM, 2016. ACM Conference on Recommender Systems (RecSys). pp. 199-202.

A. List of surveyed entities

Entity	URL	Type
Advaneo	https://www.advaneo-datamarketplace.de/en/ https://www.crunchbase.com/organization/advaneo-gmbh	DM
Airbloc	https://airbloc.org/	PIMS+DME
Atoka	https://atoka.io/ https://spaziodati.eu/	DP
AWS Marketplace - Data Exchange	https://aws.amazon.com/marketplace/ https://aws.amazon.com/data-exchange/	DM
BattleFin	https://www.battlefin.com/	DM
BookYourData	https://www.bookyourdata.com/	DP
Carto	https://carto.com/	Digital Service Providers
Caruso Dataplace	https://www.caruso-dataplace.com/ https://www.linkedin.com/company/carusodataplace/	DM
citizenme	https://www.citizenme.com/	PIMS+Surveys
Cognite*	https://www.cognite.com/	DMS
Cybernetica	https://cyber.ee/	DME
Data Intelligence Hub	https://dih.telekom.net/en/	DM
Data Republic	https://www.datarepublic.com/	DME
Databroker	https://databroker.global/	DM
DataPace	https://www.datapace.io/	DM
Datarade	https://datarade.ai/	Data Aggregator
DataScouts	https://datascouts.eu/	DP
Datasift	https://datasift.com/	Digital Service Providers
DataWallet	https://datawallet.com/	PIMS+DM
Datum	https://datum.org/	PIMS+DM
Dawex	https://www.dawex.com/en/	DM
DefinedCrowd	https://www.definedcrowd.com/ https://www.neevo.ai/	DP
Digi.me	https://digi.me/	PIMS+DME
ErnieApp	https://ernieapp.com/	PIMS+Surveys
Factual	https://www.factual.com/	Digital Service Providers
Fysical	https://fysical.org/ https://fysical.com/	DP

Entity	URL	Type
GeoDB	https://geodb.com/en/	PIMS+DM
Handshakes	https://www.handshakes.com.sg/data.html	DP
HAT	https://www.hubofallthings.com/	PIMS+DME
ifeelgoods*	https://www.ifeelgoods.com/	Digital Reward
IOTA	https://www.iota.org/ https://data.iota.org/#/	DM & DME
Lotame	https://www.lotame.com/	DP
Meeco	https://www.meeco.me/	PIMS+DME
mydex	https://mydex.org/	PIMS+DM
Ocean Protocol	https://oceanprotocol.com/	DME
Old Databroker DAO	https://www.crunchbase.com/organization/data broker-dao	
OpenCorporates	https://opencorporates.com/	DP
Openprise	https://www.openprisetech.com/	DP
Otonomo	https://otonomo.io/platform/	DM
OwnYourInfo*	http://www.ownyourinfo.com/	Personal infrastructure provider
People.io	http://people.io/	PIMS+Surveys
Qiy Foundation - DigitalMe	https://www.qiyfoundation.org/ https://digital-me.nl/	DME
Quexopa	https://quexopa.io/	DP
Refinitiv	https://www.refinitiv.com/	DP
SayMine	https://saymine.com/	PIMS
Snowflake	https://www.snowflake.com/	DME
Streamr	https://streamr.network/	DM
TelephoneLits.Biz*	https://www.telephonelists.biz/	DP
USA Sales Leads*	sales-lead.org	DP
Veracity	https://www.veracity.com/	DM
Webhose.io	https://webhose.io/	DP

Entity	URL	Type
Weople	https://weople.space/en/ https://www.hoda.digital/ https://www.linkedin.com/company/hodasrl/	PIMS+DM
Wibson	https://wibson.org/	PIMS+DM
Xignite	https://www.xignite.com/	DP

Table 9. List of surveyed entities

Entities marked with an asterisk (*) were not finally considered in the survey, due to different reasons:

- OwnYourInfo is a personal infrastructure provider, but is agnostic to the data stored in personal vaults.
- TelephoneLits.Biz and USA Sales Leads are similar to BookYourData and were discarded to avoid biasing the sample of data providers to a very specific business model.
- iFeelGoods was found not to be a data marketplace nor a PIMS. It is a company that enables an ecosystem where brands and agencies can offer their customers rewards for marketing purposes.
- Cognite was found to be a data management system, and not a data trading platform.