DELIVERABLE D1.1

PIMCity requirements and specifications

H2020-EU-2.1.1: PIMCity

Project No. 871370

Start date of project: 01-12-2019

Duration: 33 months

Revision: 02

Date: 25/10/2021

Document Information

Document Name: PIMCity requirements and specifications

WP1 – Title: Platform specification and demonstration

Task 1.1

Revision: 02

Revision Date:

Author: Roberto González Sánchez

Dissemination Level

Project co-funded by the EC within the H2020 Programme		
PU	Public	
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)	**X**

(Tick the corresponding dissemination level of the deliverable according to Annex I).

Approvals

	Name	Entity	Date	Visa
Author	**Roberto González**	**NEC**	**25/10/2020**	
Author	**Roberto Bifulco**	**NEC**	**30/05/2020**	
Author	**Daniel Oñoro**	**NEC**	**10/05/2020**	
WP3 Leader	**Rubén Cuevas**	**UC3M**	**30/05/2020**	
Author	**Angel Cuevas**	**UC3M**	**10/05/2020**	
Author	**Francisco Valera**	**UC3M**	**10/05/2020**	
Autor	**Pedro Reveriego**	**UC3M**	**10/05/2020**	
Author	**Nikolaos Laoutaris**	**IMDEA**	**30/05/2020**	
Author	**Martino Trevisan**	**POLITO**	**10/05/2020**	
WP2 Leader	**Martino Trevisan**	**POLITO**	**30/05/2020**	
WP1 Leader	**Roberto González**	**NEC**	**30/05/2020**	
Coordinator	**Roberto González**	**NEC**	**10/05/2020**	
Project Coordinator	**Marco Mellia**	**POLITO**	**10/05/2020**	
WP4 Leader	**Nicolas Kourtellis**	**TID**	**30/05/2020**	

Author	**Kleomenis Katevas**	**TID**	**10/05/2020**	
Author	**Gugliemo Bondioni**	**FW**	**30/05/2020**	
Author	**Miguel Perez**	**AUI**	**25/10/2020**	
WP7 Team	**Alessandro Bruni, Aleksandra Kuczerawy,** **Viltė Kristina Steponėnaitė**	**KUL**	**30/05/2020**	
Author	**Stefano Traverso**	**ECS**	**30/05/2020**	
Author	**Nicolás Ayala**	**GDATA**	**30/05/2020**	
Author	**Rodrigo Irarrazaval**	**GDATA**	**30/05/2020**	
Author	**Javier Calvo**	**GDATA**	**30/05/2020**	

Document history

Revision	Date	Modification
Version 1	30-05-2020	V1
Version 2	25-10-2021	The Deliverable has been revised - after the recommendations from EC experts - as follows: • Perform a review of the state of the art in users' perception on privacy. • Extended the questionnaire description and aims • Avoided the usage of "survey" and use the term "questionnaire" to avoid confusion. • Included the original data in an appendix for completeness

List of abbreviations and acronyms

Abbreviation	Meaning
PIMS	Personal Information Manager System
PDK	PIMS Development Kit
TT	Transparency Tags
PDA	Personal Data Avatar
UD	User Dashboard
P-DS	Personal Data Safe
P-PM	Personal Privacy Metrics
P-CM	Personal Consent Manager
P-PPA	Personal Privacy Preserving Analytics
D-VT	Data Valuation Tools
D-TE	Data Trading Engine
DKE	Data knowledge Extraction

Executive Summary

Deliverable 1.1 sets the requirements for the PIMCity Project, funded from the Horizon 2020 Programme (ICT-13-2018-2019) under Gran Agreement number 871370. PIMCity aims to assist companies running Personal Information Managing Systems (PIMS) to improve their operation, and to develop new business models quickly without the need of developing from the scratch the whole software stack. Moreover, PIMCity will design and develop EasyPIMS, a new PIMS specially tailored for the users of Telco companies.

Requirements have been defined by the members of the consortium developing SMOOTH. The Asociación de Usuarios de Internet have asked real users about their conception of privacy and the PIMS model. Katholieke Universiteit Leuven, as legal partner, identified the key GDPR and e-Privacy requirements applicable to PIMS. Gran Data and Cliqz, as representation of the PIMS, have stablished the requirements PIMS have for the PIMCity solutions. Moreover, Telefonica and FastWeb have proposed the requirements of the Telco companies. Finally, Politecnico di Torino, NEC Laboratories Europe GmbH, Universidad Carlos III de Madrid, Ermes CyberSecurity and LSTECH defined the technical requirements for the advanced technologies that will form the PIMCity solutions.

Requirements listed in Deliverable 1.1 form the basis of the PIMCity Project. As the work for the Project evolves, they may be subject to amendments. Deliverable 1.2 will be used to update this document. Based on the findings of the project, requirements may be amended, supplemented or deleted.

Index

1.- Introduction

1.1.- PIMCity: Background and objectives

In today's data-driven economy, the amount of data a company holds has a direct and non-trivial contribution to its overall market valuation. Data is catalysing not only business, but also governance and everyday life, across sectors, regions, time scales, economic and political systems around the world. Online advertising and marketing have been driving developments in the area by changing a decades-old industry and creating some of the biggest companies and controversies of our times in the process. Indeed, the online advertising industry is breaking records year after year in terms of both growth and total value. For instance, Zenith Media predicts a constant growth of 5% on a yearly basis, with the total expenditure on mobile advertising alone topping to 187 billion USD in 2020.

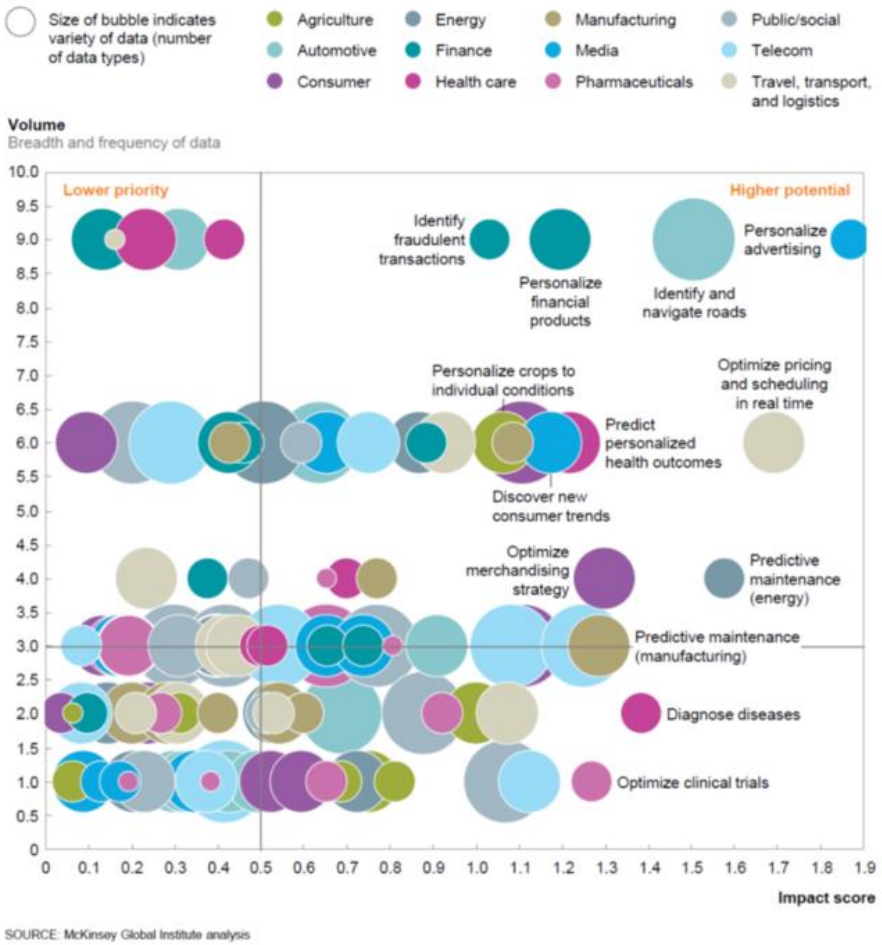

Figure 1: Data Driven decision making areas

Online advertising is just the tip of the iceberg though. Data is sought and offered across a huge spectrum of applications, promising new levels of efficiency for existing products and services as well as new capabilities just unimaginable until recently. Figure 1 depicts a summary of main areas and activities in which data driven decision making is having a pronounced impact according to a 2016 large-scale study by McKinsey. Even under the most conservative estimations, the figures for the potential of data-driven decision making

are staggering. Mobility alone could see benefits of close to 2.5 trillion dollars by 2025, whereas multiple other sectors have annual benefits in the hundreds of billions (e.g., banking 260 billion per year).

By some aspects, however, this economy is primitive: the source of value - or, so to speak, of the raw material - is often the users of online systems, and they have almost no choice but to give away their goods (their data) to very few all-powerful companies, against which they have no negotiation power. Eventually, in exchange for their goods, users get a set of services, some of which today are essential for everybody's digital life: search, map, messaging, connecting with other people, shopping. In each field, there is one clear monopoly (Google, Facebook, Amazon, to name a few). Consequently, users cannot really opt out of the deal and have to keep giving away their goods while being unable to negotiate any compensation for them. This is not a market. It resembles more the colonial economy, where peasants had no choice but to work for their colonists, with no negotiation power whatsoever. Notice that the same situation extends to companies that today have little choice and control on the information their employees (involuntarily) give away on the Internet. This raises clearly both strategical and security issues, with no easy solution.

This situation has spurred intense debates around various aspects including data protection, discrimination and bias; manipulation of public opinion and spread of fake news; competition and the creation of monopolies, automation and its impact on unemployment and economic disparity. Regulation is rushing to establish basic guidelines and barriers to contain negative consequences before it is too late. Data protection has been the focal point of activity in this regard with EU's GDPR followed by developments in other parts of the world. In a relative short amount of time it has become generally acknowledged that (loss of) privacy is not something to be left to chance and uncontrolled market dynamics.

	Potential impact: 2011 research	Value captured %
Location-based data	▪ $100 billion+ revenues for service providers ▪ Up to $700 billion value to end users	50–60
US retail[1]	▪ 60%+ increase in net margin ▪ 0.5–1.0% annual productivity growth	30–40
Manufacturing[2]	▪ Up to 50% lower product development cost ▪ Up to 25% lower operating cost ▪ Up to 30% gross margin increase	20–30
EU public sector[3]	▪ ~€250 billion value per year ▪ ~0.5% annual productivity growth	10–20
US health care	▪ $300 billion value per year ▪ ~0.7% annual productivity growth	10–20

1 Similar observations hold true for the EU retail sector.
2 Manufacturing levers divided by functional application.
3 Similar observations hold true for other high-income country governments.

Figure 2: Missing value from data-driven decision making

Beyond triggering regulatory interventions, the above practices have fuelled a direct arms race with some users who have started to adopt systems to protect their online privacy, e.g., tracker and advertising blockers. In response, services have attempted to bypass blocking using a variety of elaborate tracking techniques, and publishers have developed blocker detection technologies that redirect users to pay-walls. A continuing arms race is detrimental to the positive potential of data driven decision making. For this, McKinsey has identified only a small percentage of the potential of data-driven decision making had been realised by 2016 (Figure 2) compared to their 2011 projections.

The above suggests that although there is a huge potential from data-driven decision making, this potential will not be fully realised unless we manage to come up with the credible solutions to the mentioned controversies. Regulations and policies have made the initial steps, but they require technology to come along and offer citizens and businesses with meaningful ways to exert control over they data, just like they can control other resources they own, instead of fuelling yet another round of an arms race.

Several technological solutions and business models have appeared lately for reconciling the abovementioned tensions following suggestions and opinions that are maturing at the European policy scene and its instruments such as the EDPS. Among them, personal information management systems (PIMS), a.k.a. personal data banks or personal data vaults (we will use the term PIMS hereafter) appear to be a promising alternative to the uncontrolled collection, processing, and exploitation of people's data, including personal and sensitive ones. At a high level, a PIMS can be thought of as a software layer between end users and services, responsible for releasing data from the former towards the latter,

in a controlled manner. Different types of PIMS exist, and there is no standard definition of what a PIMS is, but several of those include capabilities like: fine grained consent management for the release of personal data towards services, the ability to revoke permissions and data, the ability to negotiate and receive payments for the release of data, privacy preserving release of aggregate analytics or raw data, dashboards for extracting knowledge and quantifications from one's own data, ability to migrate data among services, etc. Currently, PIMS from academia and industry are starting to appear and trying out various business models, with technological solutions, and go-to-market strategies in an attempt to re-write the rules of information economics on the web. Despite the large number of such attempts, none of them has yet reached business or technological maturity nor managed to attract a sizable user base.

A rushed conclusion from the above might be that the paradigm of PIMS may not be as promising as initially believed. Such a conclusion, though, is probably short-sighted, if not downright unfair and naive. Paradigm shifts that change the economy and society at large do not occur in the same circles and time scales that people expect, for example, a technological startup to succeed. Using "data as currency" and changing the current "privacy vs. utility" trade-off will take time to mature. Considering the "Data is the New Oil" metaphor, it took a more than a century for oil to become the new centre of economy, going through "wild west" economics and business practices, harsh monopolies, war, financial crashes, and controversies about its utility vs. penalty for the environment. The data economy and the new revolution that it brings will hopefully develop much faster due to the fast pace at which ICT innovation happens. Still, years would be needed, as happened for instance with cloud computing or the rise of mobile computing. An alternative business and technological model for the data economy is required, and the signs are all there. The PIMS approach is a prime candidate with many reasons supporting its adequacy. For PIMS to mature, though, technology, business model, go-to-market strategy, etc. will have to mature. Therefore, if the PIMS revolution is to succeed, it will require agility and efficiency in trying new PIMS approaches and business models.

Therein lies a major shortcoming of existing PIMS, which are monolithic in both their technology and business. Currently, a founder, or a consortium of research partners conceives a PIMS business model and technology and goes about building and developing it from scratch, and specifically targeted to that specific goal. Each such attempt requires investing a substantial amount of resources and time into a point solution in terms of a huge and largely unexplored space in terms of value proposition, technology, go-to-market strategy, modulated at the same time by multiple external factors outside the controlled of the inventors/founders of a PIMS. Such an approach is both costly and slow. Chances of success are always slim, whereas the penalty for failure is the loss of a substantial amount of effort and time. As previously said, for the PIMS approach to succeed, trial and error cycles will need to be cut shorter and made cheaper so that we can experiment more intensely and efficiently.

Aware of all this, the goal of PIMCity is to design, build, validate, demonstrate, and exploit a set of reusable, flexible, open, easy-to-use components in the form of a PIMS Development Kit (henceforth PDK, effectively an SDK for PIMS). The PDK will make building new PIMS - and extending existing ones - easier, faster, and cheaper thanks to open API. By doing so, we can help accelerate the developments towards finding the right PIMS for unblocking a fair and safe data economy. Our library will provide components for common functions and challenges confronting most PIMS such as: consent management,

data representation and management, data privacy, access control and revocation, privacy protection and privacy preserving analytics among others. Our components will be reusable, built on existing widely deployed technologies, and will have simple and clear interfaces that will facilitate building PIMS from scratch, or seamlessly integrating with the existing ones. The PDK will help in opening up the data trade market to also small and mid-sized companies as opposed to leaving it to the hands of few large internet service giants as is the case today. We will demonstrate how PDK can help small and large companies innovate in terms of technology and business models around data markets, ranging from small existing PIMS fighting to add features and get users, to specialised advertising companies and large Telcos that already hold lots of data but lack the technology to bring them to market is a sustainable and privacy compliant manner.

We strongly believe that a virtuous circle must be triggered to modify the de-facto status of data monetization. New ideas must first and foremost have answers on how to bootstrap. How do you attract the initial user base to make the platform useful to users and appealing to information buyers? How do you populate the platform with user data without pushing the burden on users themselves who are unlikely to spend time configuring and feeding with data a platform whose value they cannot yet asses? How do you provide true value for users thus converting them into advocates of your platform and using their referrals to grow your base beyond the initial critical mass? How do you achieve all the above in an informed and transparent manner that attracts users, technology companies, advertisers, and data providers into the ecosystem without jeopardising compliance with data protection laws?

For all this to happen, we firmly believe a strong catalyst is highly needed. Offering open, easy to reuse, interoperable components is an essential first step to create a market and make it open. To reach critical mass, we believe trusted, large companies must lead the way. For this, we believe Telco providers are among the best options to trigger this fundamental change. They have extensive customer bases. But they are stifled in this by huge “over the top” service providers that dominate the data market. They are strongly regulated, therefore more trustable than over the top services, and have an incentive to monetize the data they already have, but as of now have not yet been able to do so.

Finally, we are also strongly convinced that an open market for data will only flourish if we stop the arms race. For this, we must involve advertisers and end users in the whole process. PIMCity has gathered representatives from all the above-mentioned stakeholders in its consortium.

In PIMCity, we commit to design a new fair and safe data economy, implementing and releasing TRL 7 prototypes of its components, and demonstrating how it works with real end users, advertisers and operators in the area. Our objectives can be shortly summarized as follows:

Technological Objectives:

1. We will design, implement, demonstrate, and release the PIMS Development Kit (PDK), containing easy to use, interoperable, and portable software components that can be used for building new PIMS, or extending existing ones, quickly and inexpensively. Our components will cover common PIMS needs, i.e., consent management, data storage and management, data privacy, access control and revocation, privacy protection and privacy preserving analytics among others. Open API will allow access to components in a seamless manner.

2. We will build EasyPIMS, a fully-fledged PIMS for controlling, visualising, releasing, and monetizing web and mobility data, and demonstrate how easy it is to combine off-the-shelf components from the PDK with a limited amount of ad hoc code to create fast and economically powerful real world PIMS.

3. We will integrate components from the PDK to the CLIQZ and GDATA PIMS (already available in the consortium), and demonstrate how easy it is to extend existing platforms with new functionalities from our PDK.

Business Development Objectives:

4. Building on the technology produced within the project, CLIQZ and GDATA expect to reach user bases of several hundred thousands of users in three years with 5x annual growing rate. Such growth will attract the interest of data buyers involved whose number is expected to grow by a factor 10 in three years. In addition, we will bootstrap EasyPIMS by automatically importing user data from Telco providers, without requiring user intervention, mental effort, or time for anything but granting consent. For this, we will also carefully craft promoting campaigns, potentially based on monetary incentives via gift cards.

5. We will show how to further grow the user base of EasyPIMS by eliciting the help of its own users who will become the platform's best advocates having found in it a wealth of valuable services including: enhanced personal data protection in compliance with GDPR; economic benefits through fair payments for use of their data by online services; "quantify self" analytics and intuitive dashboards that will use raw browsing, calling, and mobility data to inform users about their habits.

6. We will show how to build PIMS technology and business models that stay clear of privacy-related arms races by including and providing requirements from the beginning for all the stakeholders of the personal data monetisation ecosystem, including the end users (represented by AUI), the information collectors (represented by FW and Telefonica), the advertising sector (represented by IAB affiliates), and the technology providers at the platform (represented by ERMES and GDATA), and machine learning/analytics services (represented by NEC, LSTECH, IMDEA, POLITO, UC3M, CLIQZ), in perfect accordance to regulations and data protection best-practice (granted and checked by KUL).

Broader Societal Objectives:

7. Improve citizens' trust by enabling transparency and control over several platforms, using open and interoperable modules.

8. Allow for better value creation from personal and proprietary data - enabling small and mid-size players to use the technologies developed within the project.

9. Accelerate and showcase important new ways to open up and realise a transparent data market for European citizens and industry.

1.2.- Objectives of Deliverable 1.1

Deliverable D1.1 gathers and compiles the requirements from the different stakeholders to ensure the success of the PIMCity project. To this end, we do not limit ourselves to those requirements that ensure the technical operation of the PDK and the EasyPIMS platform.

Instead we involve all the actors in the data value chain to ensure a comprehensive list of requirements. It includes:

- The legal requirements under the European legislation (with a special focus on the GDPR and the ePrivacy laws) to ensure a smooth operation during the execution of the project, and in its subsequent exploitation. This document presents a brief overview of some key and potentially relevant legal frameworks and requirements that will be further refined and revealed in detail in the D7.2.
- The needs exposed by the final users in order to increase the trust on the data platforms and the PIMS in particular. This document summarizes the results of a focus groups organized with tens real Internet users in order to understand the drivers and obstacles they identify in the trading/sharing of their data to third parties. We settle the requirements and specification to put the users in the centre of the data economy, by presenting the different challenges as perceived by end users that need to be solved by the different PIMCity components.
- The requirements of the different PIMS models. Since the current ecosystem of PIMS includes companies with very different business models, and therefore, very different needs, we include in this document a comprehensive list of requirements, both general for all the PIMS and specific to their different business models.
- Last, but not least, we presented the technical requirements. Using the expertise of the different technical partners participating in PIMCity we set a realistic list of requirements for both the PDK in general, and each tool in particular. For this, we provide a set of functional and non-functional requirements that our solution should fulfil.

This document states a preliminary version of the requirements of the EasyPIMS that will be revised in the following D1.2 At last, this document is concluded with a summary of the lessons learnt during the collection of the requirements.

1.3.- Changes from Version 1

This document updates the Version 1 and modify the *Section 2. Final Users requirements for the PDK* and adds a new *ANEX*. In the Section 2 it includes a review of the state of the art about user preferences on privacy, the annex reflects the results of the questionnaire done during the PIMCity project.

2.- Final Users requirements for the PDK

One of the objectives of PIMCity is to ensure that individuals, the end users, can control who uses their personal data, stipulate for what purposes it may be used, give informed consent, know the value of this data in the marketplace and obtain compensation or benefits for its use.

In order to establish these requirements, from the user's perspective, we have developed a study with two working groups prior to the preparation of these specifications and have subsequently carried out a questionnaire from which 243 valid responses have been received.

The objective of this previous work with focus groups has been to identify those aspects that most concern the users in order to be able to transfer them to the specifications of the

different modules that are going to be developed in the project and that must be taken into account both when developing the PDK components and when implementing the EasyPIMS solution.

2.1.- Users' perceptions of privacy

One of the aspects to consider is to know the perception of the users about their privacy

To understand the perception of privacy, we checked the state of the art, and after, we did our own questionnaire to get more insights. We have analyzed in detail other surveys, studies and papers already published and we analyze the responses received from the questionnaire.

The objective of this review is to build a picture of consumer attitudes in general and to identify broad themes in terms of problems and concerns.

To organize our analysis, we considered a series of high-level questions for each of the topics to better explore the diversity of issues addressed in the consumer surveys and academic research papers.

Here below the topics and the questions presented in the survey:

A) Knowledge and understanding about data processing.

How much do users know, or think they know, about their personal data?

How do they react to - and understand - the terms and conditions contained in privacy policies?

B) Users' control over their personal data

Do users believe they control their data and to what extent?

Do they engage in control of their data?

C) Attitudes towards data processing and personalized advertising.

What are users' attitudes towards data processing?

What are users' attitudes towards personalized advertising?

How do users perceive the benefits and harms of data processing in relation to personalized advertising?

2.1.1.- Users knowledge and understanding about data processing

How much do users know, or think they know, about their personal data?

People care about privacy. Across continents, age, gender, and levels of education, people overwhelmingly think privacy is important (Telefonica - IE , 2020).

Users understand that personal data is valuable to them, and they overwhelmingly agree that companies and Internet platforms are the ones that benefit most from the processing of their data, and that these companies and platforms collect and process it for profit (Ipsos Mori, 2016) (DMA-Acxiom, 2018)

Most users are not sure which data do the platforms have about them, but there is greater awareness of the information they voluntarily provide, compared to the amount of data passively collected (Harris Interactive, 2019) (Doteveryone, 2018)

There is a common perception that platforms collect a large amount of data, yet only few users are aware of the true volume of data collected about them. (Ipsos Mori, 2016)

Both academic research and surveys agree that most users have a very basic understanding of how their personal data is processed. In particular:

- There is a greater recognition of "active" methods of data collection over “passive" methods, although most users say they are aware of cookies. (Ofcom, 2019, pág. Table 97)

- Users are more aware of the evident uses made of their data (e.g. advertising or personalized recommendations) than of the hidden uses (e.g. price discrimination according to profiles). (Doteveryone, 2018)

- Few users are aware of the extent to which data are shared, or that data may be combined to form profiles before being shared. (Which, 2018)

- There is also evidence that, over the time, is increasing the awareness of how data are collected, used, and shared. (Information Commissioner's Office, 2019)

How do they react to the terms and conditions contained in privacy policies and how much do they understand?

Only a small minority of users say that they always read privacy policies or terms and conditions of use. (Information Commissioner's Office, 2019)

On the other hand , academic research shows that the number of users who read online policies in practice is much lower than indicated in consumer surveys by the users themselves. (Harris Interactive, 2019)

Research conducted in 2007 estimated that a user would have to spend several weeks a year reading the policies of every website they visit, and since 2007 the length of the associated text has continued to grow. (Into The Minds, 2018) (EC Eurobarometer The General Data Protection Regulation, 2019) (EC Eurobarometer The General Data Protection Regulation, 2019) (MacDonald, 2008) .

Approximately half of users report not understanding online policies when they read them (Harris Interactive, 2019).

Users point to "legal terms" and "general statements" to justify difficulty in reading them and studies have shown that understanding privacy policies requires a high level of reading proficiency (Cardogan, 2004).

2.1.2.- Users' control over their personal data

There is consensus that few users feel they have control over their data (Ipsos Mori, 2016) (EC Eurobarometer The General Data Protection Regulation, 2019). Some users feel they can manage some actions (such as deciding whether or not to enter certain information in a form or visit a website), but they feel to have little control over how their data is used or

shared once they are using an online platform, application or service (Ipsos Mori, 2016). In addition, as users learn more about data processing, they begin to feel they have less control and less confidence in their ability to manage some aspects of their data processing. (Which, 2018)

Users also note that it is difficult to engage with companies that collect and use their data because they feel that:

- Are disempowered by their lack of knowledge and transparency about how companies collect, use and share their data
- It is difficult to access and modify the personal information held by companies
- They are dependent on data-driven services that they do not believe they can opt out of
- There is a perceived lack of alternatives if they want to stop using particular service, app or platform

(Doteveryone, 2018) (Which, 2018) (EC Eurobarometer The General Data Protection Regulation, 2019)

Among users, there are discrepancies concerning their ability to set and control the privacy features of their browser and social network accounts. Generally, users report to be confident to set and control privacy features, but when they are asked to do it in a workshop or focus group, many users have difficulty accomplishing these actions (Ipsos Mori, 2016) (The Centre for Data Ethics and Innovation, 2020). Participants who reported a problem, commonly say that privacy settings were often complicated to find and to be used.

It is unclear how often users change their privacy settings, although some studies note that the majority of users report that they have changed their privacy settings at least once (of their browser or their social network account) (EC Eurobarometer The General Data Protection Regulation, 2019) (EU Eurobarometer e-Privacy) . However, few users report that they find easy to access and change the personal information held by a company or find out how data is collected, stored, accessed and changed (Information Commissioner's Office, 2019). Perhaps due to this, most users agree on that default settings should stop their information from being shared (EU Eurobarometer e-Privacy) (Illuminas for Citizens Advice, 2016).

As for the General Data Protection Regulation (GDPR), most users have heard about it, although only half of them have some understanding of what the GDPR entails (EC Eurobarometer The General Data Protection Regulation, 2019). In general:

- The most widely known right was the right to access the user's own data
- The most exercised right is the right to object to receiving direct marketing
- The least known and least exercised right is the right to have a say when decisions are automated.

2.1.3.- Attitudes towards data processing and personalized advertising.

It is clear from the surveys and academic research that users report that privacy is important to them, but it is hard to determine exactly how many users value their privacy. Research has also reported what has been called the 'privacy paradox': users say privacy is very important to them, but their actions and behaviours indicate otherwise.

What are users' attitudes towards data processing?

Most users consider that data processing is inevitable in modern life and that it will happen even more frequently. Despite this, there is evidence that users do not fully understand the role of their data processing: only the most informed understand that it is the "price" they pay to have free access to products or services online (Which, 2018). Qualitative surveys have also revealed that data processing may not be a priority concern for most users when using the Internet (Doteveryone, 2018) (Which, 2018).

This does not mean that users are comfortable with data processing. On the contrary, there is evidence that the majority of users are uncomfortable with data processing or are concerned about their privacy (Ipsos Mori, 2016) (EC Eurobarometer: Online platforms, 2016) .

Most users are concerned about their data being shared, and this concern increases when data sharing is perceived to be done without the consumer's consent (Ipsos Mori, 2016). There is also evidence that the more users learn about data processing, the more concerned they become about it (Harris Interactive, 2019) (Which, 2018).

Surveys suggest that users are more comfortable with data processing and they accept it when:

- The data are considered relevant (e.g., location data on maps) and are not considered sensitive (e.g., age vs. household income) (DMA-Acxiom, 2018)
- The data are anonymized and aggregated (Royal Statistical Society, 2017)
- The use of the data has a clear benefit to the consumer or society
- The government agencies process their data rather than commercial third parties (DMA-Acxiom, 2018)

Only a minority of users trust the processing of their data by online platforms, and among these, Social Networks are the least trusted by users. People tend to feel that they cannot trust companies and institutions to protect their privacy and use their personal data in responsible ways. From the Big Tech companies, Facebook is thought to be the most untrustworthy, and Apple and Amazon the most trustworthy (Telefonica - IE , 2020)

Some users also believe that platforms will do what they want with their data regardless of what the consumer agrees to (Doteveryone, 2018). This is important considering that most users say that trust is one of the most important considerations for them when making decisions in the online environment (Open Data Institute , 2019).

How do users perceive the benefits and harms of data processing in relation to personalized advertising?

In the case of personalized advertising, users recognize that it increases the relevance of what they are shown (Which, 2018). However, there is also evidence that very few users are willing to share their data in exchange for these benefits:

- This survey shows that only 15% of respondents were happy for online companies to collect and use their data in exchange for a personalized service (Ofcom, 2019).

- This one shows that only 5% of respondents felt that they benefited from companies using their personal information to send them personalized ads (Ipsos Mori, 2016) .

On the other hand, users find it difficult to point to specific examples of harm as a result of data processing or targeted behavioral advertising. However, a number of general concerns do emerge from these surveys, including

- Loss of privacy

- Use of inaccurate or personal data in automated decisions

- Loss of control over data and ads

- Data security breaches

- Lack of trust in organizations that enable data processing and targeted advertising.

One of the reasons users may have difficulty articulating the benefits or harms of personalized advertising is the fact that - as noted above - few users understand what, how and why data is collected and shared, or how targeted behavioral advertising works. This, coupled with the inherent opacity of data processing and online targeting, and psychological biases, indicates that users' ability to anticipate potential harms and benefits is likely to be very limited.

Which?'s qualitative study examined the methods used to collect data for personalized advertising. Although Which? found that most participants preferred to receive targeted, rather than generic, advertisements, participants also had a clear preference for having to opt in to data collection for targeted advertising, rather than opting out (Which, 2018).

2.2.- Design Principles: User centred model

One of the aspects that has emerged repeatedly in the working sessions has been the lack of knowledge of how user data is collected (lack of transparency), the lack of confidence in the large data aggregators (social networks, search engines, operating systems) and the little control that users have over those who currently handle their personal data.

When asked to choose which of the three models shown in the figure (APIs for each service provided, aggregation model or User Centred Model), users mostly prefer the "user centric" model.

PIMCity is aligned from its conception with the approach promoted by the MyData movement[1] which advocates a paradigm shift in the management and processing of personal data to move from a model focused on companies that collect data (with little transparency and very little control) to a system focused on the person and totally transparent.

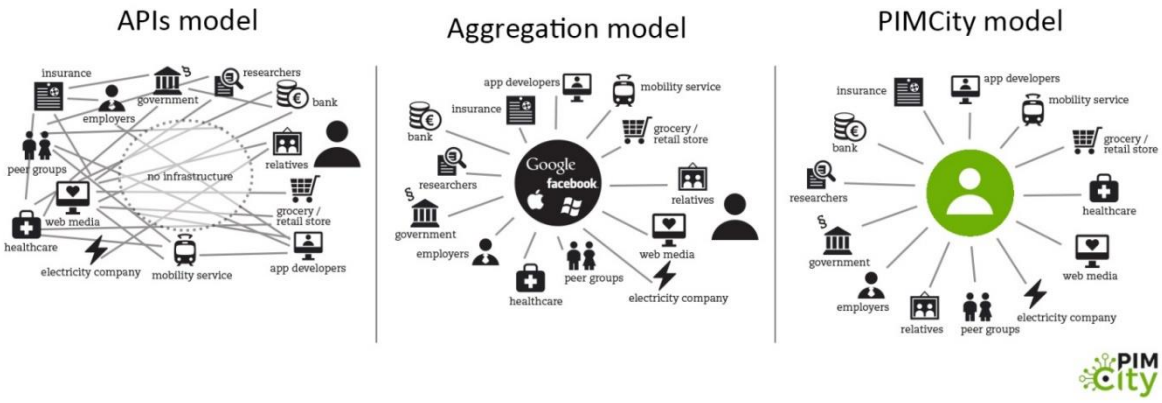

Figure 3: Data models comparison

Another aspect pointed out by the working groups is the complexity in configuring the different privacy options and the difficulty in understanding the terms and conditions of use proposed by the current applications and services that are used, which leads to accepting them without reading them but with a degree of mistrust that grows as facts and news about abuses, misuses and attacks that affect the data of millions of users are produced. It is therefore essential that each user's personal data can be controlled, managed, accessed and used in a simple, secure and intuitive way, making it a valuable and reusable resource. Data that can be used to create new services and new business models and economic growth for society.

[1] http://www.mydata.org

The third element has to do with the variety of interfaces that each application offers for privacy management which becomes a new barrier for users. This leads us to suggest that PIMcity, through the tools to be developed in its PDK development kit and its EasyPIMS application example, should allow a decentralized management of personal data so that the interface for the user is always the same.

With all this information we identified that one of the design principles in PIMCity should place the user at the centre of the ecosystem, so that they are empowered actors, not passive objects who are given information and tools to manage and control the use made of their data and therefore their privacy. Figure 3: Data models comparison

2.3.- Consent management

The common "I have read and agree the terms of use" acceptance mechanism is not appropriate, because the terms of service and privacy policies are too long and complicated to understand, which has led companies to take advantage of it by increasingly demanding access to an even greater variety of personal data in exchange for better services.

Currently users give their legal consent for the collection and exploitation of their personal data by clicking on the box confirming that they have read and accepted the terms of use of each application or online service that they use. They usually confirm without reading in a very high percentage and if they do, they do not understand what they have read.

The PIMCity model is based on the fact that it is the users who have the control and information to decide how their data should be used, allowing the collection and reuse of their personal data in a way that maximizes the benefits obtained and minimizes the impact on privacy.

PIMCity through the PDK provides the tools that allow other organizations, applications and services to access, obtain and use the data sets containing personal information such as socio-demographic data, purchasing data, mobility data, health data, financial data or data derived from the different online services, ensuring that individuals remain in control of their data and can do so in a secure, easy and compliant way, in accordance with the legal requirements advocated by the EU and the proposed ethical principles.

The user in PIMCity will have a control panel (dashboard) from which you can easily control all your personal data from a unique way regardless of where they are generated and who uses them. From a single place the user decides what data they want to share, who gives them permission to access them, for what purpose and what is the benefit of using such data and all in a simple, understandable and secure.

Consent management is the main mechanism to enable and meet the legal requirements of the GDPR. Through the Dashboard, individuals can give instructions for services to obtain and process data in accordance with the consents that the individual gives to each of them individually and in an informed way.

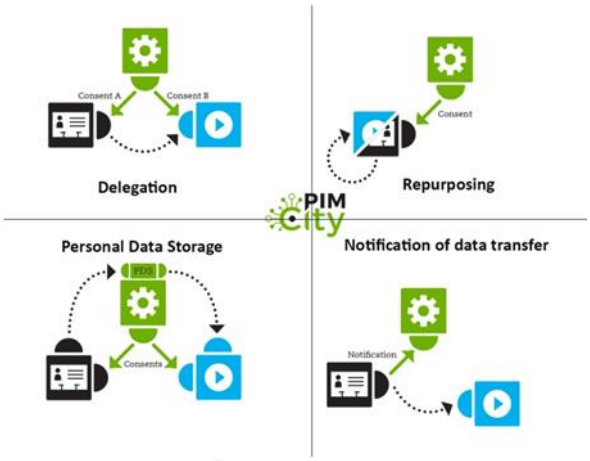

Figure 4: Consent and Data flows

In PIMCity, the consents must be dynamic, easy to understand for the users, readable by the systems to be automated, standardized and managed in a coordinated way by the user through its Dashboard.

Consents can be delegated so that the person generating the data provides it to the person who has requested it for a specific use (Delegation), consent can be requested to use data that is already being used and generated for a different purpose (Repurposing), in other cases the data will be directly stored in the user's system (Personal Data Storage) and there may be cases in which an administrative body has the legal possibility of transferring data to a third party under the premise of informing the user of this situation (Notification).

2.4.- Interoperability

In the PIMCity architecture, data flows from the source that generates it to the service or application that uses it as long as it is authorized by the user. It is therefore important to clarify the flow of consents or permissions which is where the control lies and which may be different from the actual flow of data and its storage.

The user, through the dashboard can, control, inform himself and give consent for the use of different data for multiple services and applications.

The architecture of PIMCity must allow users to integrate new data sources, and connect them to new services that is to say that there is interoperability. This property is an important element to generate confidence in the PIMCity system. Interoperability is the main advantage offered by the proposed solution PIMCity and at the same time is the main challenge because it requires a process of standardization and standardization of consent mechanisms, formats and semantics.

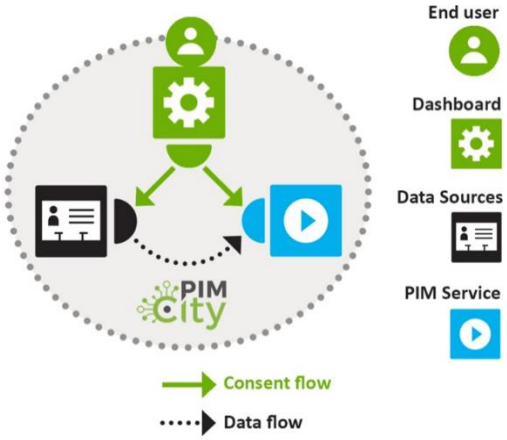

Figure 5: PIMCity consent flow

2.5.- The value of the data

An important contribution of PIMCity are the tools that, on the one hand, provide transparency over the market price for a set of data and, on the other hand, facilitate the contact between those who need that data and agree on a price in a simple, safe and non-invasive way from the perspective of users.

One of PIMCity's contributions is that it provides information to the user about the value that a data or a set of data has at a given time in the market which will help him or her to decide whether or not to accept a proposal.

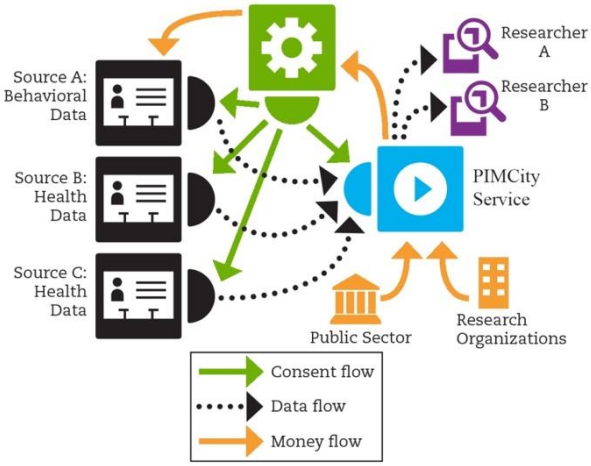

Figure 6: PIMCity money flow

2.6.- Data aggregation

One of the evidences detected in the working sessions has been that under the heading of personal data, potential uses of the same and types of consent, each person interprets different things.

It is important to mitigate this risk by carefully designing how to group data and consents so that they are understandable to individuals and so that the different modules that make up PIMCity address these elements in a homogeneous way.

When standardized they can be made readable by applications and are easier to compare, group, visualize, manage and process.

The Creative Commons licensing framework provides an example of how the equally heterogeneous realm of copyright was harmonized under a common set of standardized licenses.

One of the challenges has been to group personal data to facilitate joint information and decision making about a group of data while also allowing the user to make decisions at different granularity.

This exercise of grouping the data has allowed us to identify the following groups of data

- **Identity data**: This is the data that identifies a person biunivocally (name, contact details, identity). Identifying data are ussualy disaggregated from the rest of the data so that when they are shared with third parties the owner of the data cannot be identified. However, the user may decide to share some of these data and that some of them may be necessary for the provision of a service (receiving a postal package, a communication, etc.).

- **Sociodemographic**: Age, Gender, Family status, Education level, Income level, Occupation level, Language give general information about each person

- **Behaviour**: browsing history, internet searches, use of social networks, advertising interactions

- **Location**: Country, State, Province, City, Zip Code, Street, Number

- **Mobility**: GPS data, operator data, daily distances

- **Shopping**: Shopping

- **Finance data**: Banking movements, Card movements

- **Special categories" of personal data**: personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation

The characterization of consents should allow the user to easily answer these questions:

- The **purpose** of the processing, i.e. the use of the data requested

- If the data is **shared** on to third parties

- **Where** the data to be transferred will be stored

- **Rights** that I can exercise over the data transferred

- The **duration** of treatment for how long the data is stored

- The **benefit** or consideration obtained by the user for the transfer of his data.

2.7.- Transparency tags

Transparency tags are intended to provide key information to know on one hand the degree of reliability in terms of privacy and personal data processing.

In this matter, the objective of the working groups has been to identify the characteristics and elements that should be included so that they are intuitive and efficient. The most interesting conclusions are the following:

- ✓ They must include a visual element (color code or rating stars) that allows the degree of trust to be identified in a simple way.
- ✓ Informing about the person responsible for the data
- ✓ Informing about the purpose of the data
- ✓ Informing about the transfer to third parties

3.- Legal Frameworks

Section 3 provides overview of some key and potentially relevant legal frameworks for PIMCity project, in particular overview of the European Union (hereinafter the EU) Regulation on the protection of individuals concerning the processing of personal data and on the free movement of such data[2] (hereinafter the GDPR), e-Privacy directive[3] (hereinafter the ePD), the Regulation on a framework for the free flow of non-personal data in the EU[4] (hereinafter the NPD Regulation), the Regulation on promoting fairness and transparency for business users of online intermediation services[5] (hereinafter the P2BR), the Directive on the Security of Network and Information Systems (hereinafter as the NIS Directive)[6], the so-called Cybersecurity Act Regulation (hereinafter the Cybersecurity Act)[7], the Directive establishing the European Electronic Communication Code[8] (hereinafter EECC), the Directive on contracts for the supply of Digital Content and Services Services[9] (hereinafter as DCD), the Directive on the harmonisation of certain aspects of copyright and related rights in the information society[10] (hereinafter as the InfoSoc Directive), the Directive on copyright and related rights in the Digital Single Market[11], including but not limited. The section does not provide an exhaustive list of all of the relevant legal requirements, however. The initial overview of the key and potentially relevant legal requirements will be provided in Deliverable 7.2 (M6).

[2] Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] O.J.E.U., L119/1.
[3] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector.
[4] Regulation 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union [2018] O.J.E.U. L303/59.
[5] Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services, PE/56/2019/REV/1, OJ L 186, 11.7.2019.
[6] Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union OJ L 194, 19.7.2016.
[7] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act), PE/86/2018/REV/1, OJ L 151, 7.6.2019
[8] Directive (EU) 2018/1972 establishing the European Electronic Communications Code (Recast) OJ L 321, 17.12.2018.
[9] Directive (EU) 2019/770 of the European Parliament and of the Council of 20 May 2019 on certain aspects concerning contracts for the supply of digital content and digital services, PE/26/2019/REV/1, OJ L 136, 22.5.2019.
[10] Directive 2001/29/EC of the European Parliament and of the Council on the harmonisation of certain aspects of copyright and related rights in the information society [2001] *O.J.E.U.* L167/10.
[11] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] *O.J.E.U.* L130/92.

3.1.- Legislative Overview

3.1.1.- Privacy and Data Protection

The processing of data relating to individuals in the context of the **PIMCity** project may interfere with their right to privacy and data protection. Therefore, it is necessary to address legal frameworks which concern these rights. Section 3.1 provides an overview of the relevant legal frameworks with a particular focus on the GDPR.

Privacy and data protection are both fundamental rights, protected by largely intertwined legal frameworks of the Council of Europe and the EU, in particular by the European Convention on Human Rights[12] (hereinafter the ECHR) and the Charter of Fundamental Rights of the European Union[13] (hereinafter further as the CFREU). While the right to privacy is protected under the ECHR[14] and the CFREU[15], data protection has no direct counterpart in the ECHR but is enshrined in the CFREU.[16] The absence of a clear provision on data protection in the ECHR is compensated by two factors. First, the ECtHR has already interpreted the ECHR[17] so to encompass a right to data protection.[18] Second, a separate convention adopted in 1981 specifically addresses data protection issues in the Council of Europe's jurisdiction (Convention 108).[19]

Privacy and data protection differ in their scope. Privacy concerns private and family life, home and correspondence/communications.[20] Data protection concerns personal data which is defined as any information relating to an identified or identifiable natural person.[21] There might be situations where there is an interference with an individual's private life without any processing of personal data. The same reasoning applies *a contrario*, when there is a processing of personal data that does not have any impact on the person's private and family life. But in many cases, both aspects will enter into play. Given the above, a single interference might trigger two different assessments on a different basis of the ECHR and the CFREU.[22]

[12] Council of Europe, Convention for the Protection of Human Rights and fundamental Freedoms, signed in Rome, 4 November 1950.
[13] Charter of Fundamental Rights of the European Union, O.J.E.U., 18 December 2000, C 364/01.
[14] Art. 8 ECHR.
[15] Art. 7 CFREU.
[16] Art. 8 CFREU.
[17] Art. 8 ECHR.
[18] See, a.o. the following cases: ECtHR, Amann v. Switzerland, n. 27798/95, ECHR 2000-II, para. 65; ECtHR, Rotaru v. Romania, n. 28341/95, ECHR 2000-V, para. 43. On the broad interpretation given by the ECtHR to privacy, see Gordon Nardell Qc, 'Levelling up: Data Privacy and the European Court of Human Rights', Data Protection in a Profiled World (Springer, Dordrecht 2010) 43 <https://link.springer.com/chapter/10.1007/978-90-481-8865-9_3> accessed 14 March 2020.
[19] Council of Europe, Convention for the protection of individuals with regard to automatic processing of personal data, Strasbourg, 28 January 1981. This Convention is currently under revision; on that point, see: Cécile De Terwangne, 'The work of revision of the Council of Europe Convention 108 for the protection of individuals as regards the automatic processing of personal data' (2014) 28 International Review of Law, Computers & Technology 118-130.
[20] Art. 8 ECHR and Art. 7 CFREU.
[21] Art. 4(1) GDPR.
[22] One on the basis of Art. 8(2) ECHR and Art. 52(1) CFREU – for the right to privacy – and the other on the basis of Art. 8(2) ECHR and Art.s 8(2) and 52(1) CFREU – for the right to data protection.

Given the foreseen activities in the **PIMCity project**, it is highly likely that the processed data will qualify as personal data, and its processing will fall within the scope of application of both the CFREU and the GDPR. It is also likely that corresponding processing activities will be considered as interferences with the right to privacy. The interferences with both of the rights will have to be compatible with the legal requirements, including the overarching requirements stemming from the ECHR and the CFREU with regard to the right to privacy and the requirements stemming from the CFREU with regard to personal data protection.[23] Besides, some detailed requirements are provided in the GDPR and the ePD, addressed briefly below.

3.1.2.- General Data Protection Regulation

In order to decide whether the requirements of the GDPR are applicable to the PIMCity partners, it has to be assessed whether or not the data processed can be qualified as personal data as defined above. Further, it has to be assessed whether the requirements are relevant taking into account the material, personal and territorial scope of application.

Material, personal and territorial scope of application

With regard to material scope, the GDPR is applicable in case of processing of personal data. The GDPR applies to the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.[24] Processing means any operation or set of operation which is performed on personal data or on sets of personal data, whether or not by automated means. Processing activities include collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.[25] Basically, it encompasses everything that can be done with data, from their original collection to their definitive erasure. Given the goals of the PIMCity project, it is highly likely that that data will be processed and that the first condition for the material application of the GDPR will be fulfilled. Furthermore, for the processing of data to fall within the scope of application, the data must be personal.

The GDPR defines personal data as any information relating to an identified or identifiable natural person ('data subject'). It further adds that an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.[26]

Given the goals of the PIMCity project, it is highly likely that the data will be considered personal data in light of the potential of identifiability. The partners shall take into account that in case the personal data would be pseudonymised in a retraceable way (e.g. using correspondence lists or two-way cryptography algorithms), it shall still be considered personal data. The evaluation changes in case it would be done in a non-retraceable way (e.g. using one-way cryptography algorithms) and individuals would be no longer

[23] One on the basis of Art. 8(2) ECHR and Art. 52(1) CFREU – for the right to privacy – and the other on the basis of Art. 8(2) ECHR and Art.s 8(2) and 52(1) CFREU – for the right to data protection.
[24] Art. 2(1) GDPR.
[25] Art. 4(2) GDPR.
[26] Art. 4(1) GDPR.

identifiable. Non-retraceable pseudonymisation techniques generally create anonymised data that are not subject to data protection rules. Determining the applicability of data protection rules to data that have undergone techniques and methods to reduce identifiability will, therefore, require a case-by-case analysis of the factual circumstances surrounding the processing operations.

With regard to personal scope, it shall be noted that obligations of the particular partner may be different depending on its role in the light of the GDPR, e.g. depending on whether a partner would be considered a controller or a processor. Qualifying the role of each partner is the very first step towards effective compliance.

Controller shall be understood as the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data.[27] It is particularly important that the natural or legal person may determine the purposes and the means of the processing alone or jointly with others. In other words, situations of joint controllership are conceivable when, for a single processing operation, a number of parties jointly determine all the purposes and the means of the processing to be carried out. In that case, collaborating entities shall be jointly and equally responsible for compliance with all applicable data protection requirements. This situation must not be confused with separate controllership, where different parties do not pursue the same objectives or do not rely on identical means. In that case, each party bear its own reasonability to comply with the GDPR and shall be individually liable in case of non-compliance.

Processor shall be understood as the natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.[28] It is relatively frequent that a controller – who determines the purposes of the processing activities – assigns parts of the operations to a separate legal entity. The GDPR imposes limited obligations to the latter, mainly with regard to security, breach notifications and data transfers.

The relationship between the controller and the processor shall be governed by a contract or other legal act that is binding and that sets out the subject-matter and duration of the processing, its nature and purpose, the type of personal data and categories of data subjects and the obligations and rights of the controller.[29]

In the context of the PIMCity project, it is crucial to identify every processing operation and the role of the parties involved in the data processing chain.

With regard to personal scope, the GDPR enounces a twofold rule. First, the Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not. Second, the Regulation also applies to the processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union, where the processing activities are related to: (a) the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union; or (b) the monitoring of their behaviour as far as their behaviour takes place

[27] Art. 4(7) GDPR.
[28] Art. 4(8) GDPR.
[29] Art. 28(3) GDPR.

within the Union. To ascertain whether the offer of goods or services actually targets EU data subjects, several factors must be taken into account. In particular, specific attention should be paid to: the use of an EU language, the display of prices in EU currency, the ability to place an order in EU languages and the potential references to EU users or customers.[30] The mere accessibility of the controller's website on the EU territory should not, however, be enough.[31] Monitoring activities encompass situations where natural persons are tracked on the internet, notably through techniques which consist of profiling a natural person, particularly in order to take decisions concerning her or him or for analysing or predicting her or his personal preferences, behaviours and attitudes.[32]

General principles and rules

The GDPR provides a number of principles and rules for both controllers and processors. In brief, the principles include: (a) lawfulness, fairness and transparency, (b) purpose limitation, (c) data minimisation, (d) accuracy, (e) storage limitation, (f) integrity and confidentiality and (g) accountability[33]. The key rules include, e.g. (i) obligations for the controller to comply with all the above-mentioned principles, (ii) to implement appropriate technical and organisational measures to ensure and demonstrate that processing operations are performed in accordance with data protection rules, (iii) to comply with certain requirements with regard to their relationship with processors, (iv) to maintain a record of processing activities which shall contain, at least, the elements listed in the GDPR, (v) to ensure a level of security which is appropriate to the risk of varying likelihood and severity for the rights and freedoms of natural person raised by the processing operations[34].

Besides, the GDPR provides a number of specific data subject's rights[35]. **The right of access to certain information**, i.e. to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, access to the personal data and certain information. This information may be, e.g. the purposes of the processing, the categories of personal data concerned, the recipients or categories of recipients to whom the personal data have been or will be disclosed, the retention period, the existence of the right to rectification or erasure, and others.

The GDPR also provides **the right to rectification** which allows the data subject to obtain from the controller the rectification of inaccurate personal data concerning him or her without undue delay. Data subjects also have **the right to have incomplete personal data completed**, including by means of providing a supplementary statement. Data subjects also have **the right to erasure**, also known as the so-called right to be forgotten. This

[30] See on that point Art. 3 GDPR Rec. 23 GDPR. All these elements must be taken into account when determining whether or not the organisation targets European consumers and/or citizens.
[31] Criteria used in international private law under Regulation 1215/2012 of the European Parliament and of the Council of 12 December 2012 on jurisdiction and the recognition and enforcement of judgments in civil and commercial matters (Brussels 1 Regulation) may prove particularly useful in determining whether or not a controller is targeting EU data subjects. See a.o. on the targeting of consumers: CJEU, Pammer v. Reederei Karl Schlüter GmbH & Co KG and Hotel Aplenhof GesmbH v Olivier Heller, joined cases C-585/08 and C-144/09, para. 93-95.
[32] See also: Jiahong Chen, 'How the Best-Laid Plans Go Awry: The (Unsolved) Issues of Applicable Law in the General Data Protection Regulation' (2016) 6 International Data Privacy Law 310.
[33] Art. 5 GDPR.
[34] See Chapter IV GDPR.
[35] See Chapter III GDPR.

prerogative allows data subjects to seek the erasure of their personal data under certain conditions.

The GDPR also provides data subjects with the opportunity to **restrict the processing of their personal data**. In other words, the controller may keep the personal data at stake, but must refrain from using them during the period for which the right applies.[36] This can be requested when the accuracy of the personal data is contested as well as when the processing is unlawful and the data subject opposes the erasure of the personal data and requests the restriction of their use instead. Besides, it can be requested when the controller no longer needs the personal data for the purposes of the processing, but they are required for the establishment, exercise or defence of legal claims as well as when the data subject has objected to the processing, pending verification of the overriding grounds mentioned in the GDPR.[37]

The GDPR also provides **the right to data portability**, which allows data subjects, firstly, to receive their personal data in a structured, commonly used and machine-readable format. Secondly, it allows data subjects to transmit data to another controller without hindrance from the original controller. Certain conditions apply, however[38]. Where technically feasible, data subjects should even have the right to have the said data transmitted directly from one controller to another. When the processing is necessary for the performance of a task carried out in the public interest or is based on the controller's legitimate interests, the GDPR offers the data subject the possibility to object, on grounds relating to his or her particular situation, to the processing of their personal data. As a consequence, the controller is no longer allowed to process the said data unless it demonstrates compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject or for the establishment, exercise or defence of legal claims. When the personal data are processed for direct marketing purposes, data subjects may object at any time, without the need to conduct the above-mentioned assessment.

The GDPR also grants the data subject **the right not to be subject to a decision based solely on automated processing** which produces legal effects concerning him or her or significantly affects him or her.[39] This right is not applicable when the decision is necessary for entering into, or performance of, a contract between the data subject and a data controller, authorised by EU or Member States law or based on the data subject's explicit consent.

All of the principles, rules and data subject's rights will be taken into account when developing legal requirements for the PIMCity project in the Deliverable D7.2, which will further substantiate and contextualise these various prerogatives.

[36] Rec. 67 GDPR suggests that compliance with this provision could be achieved by temporarily moving the selected data to another processing system, making the selected personal data unavailable to users, or temporarily removing published data from a website.
[37] Art. 21(1) GDPR.
[38] Art. 20 GDPR.
[39] This would be the case when data subjects face an automatic refusal or their online credit application or e-recruiting practices without any human intervention (Rec. 71 GDPR).

3.1.3.- e-Privacy Directive

The provisions of the GDPR apply horizontally to all sectors, without any differentiation. However, within the EU legal framework, the ePD provides additional rules for the processing of personal data in the electronic communications context. E.g. there are certain rules with regard to cookies, processing particular data such as traffic or location data, security requirements, including but not limited.

The ePD obliges providers of publicly available electronic communications services to ensure, through appropriate organizational and security requirements, the privacy and confidentiality of electronic communications.[40] The obligations provided in the ePD may be relevant for the PIMCity project and will be revealed in detail in the Deliverable D7.2.

It is particularly relevant that the European Commission, given the rapid development and an ever-growing volume of data used in the electronic communications environment, is considering replacing the ePD by a new electronic privacy regulation (hereinafter the ePR). In particular, the European Commission launched its proposal for a new ePR, aiming at complementing the GDPR, to protect the right to privacy and to the confidentiality of communications in January 2017. However, there is currently no certainty with regard to the final outcome. After the lengthy legislative procedure, the Permanent Representatives Committee of the Council of the EU (COREPER) has rejected the Council's position on a draft ePR on 22 November 2019. The developments will be closely monitored and taken into account in the course of the PIMCity project.

3.2.- Platforms, Free Flow of Data and Data Market Place

3.2.1.- European Commission Communication "Building a European Data Economy"

In the last years, the EC Communication has focused part of its activity on supporting the development of the EU data economy, considered a pillar for the deployment of the EU Digital Single Market. Consequently, the EC has published its Communication on "Building a European Data Economy" (hereinafter as the Communication).[41] The Communication is the first non-binding legislative initiative in the data economy context for regulating and promoting the free-flow of non-personal data within the EU. In the Communication, the EC not only provides crucial definitions (e.g. on data market place) but also sets the scene for upcoming legislative initiatives such as the Regulation on free-flow of non-personal data.[42] The Communication and the Regulation on free-flow of non-personal data aim to pave the way for the enhancement of cooperation between different actors involved in the data market environment, increasing the economic opportunities for the actors involved.

The EC's Communication defines data market place as a market *'where digital data is exchanged as products or services derived from raw data – on the economy as a whole. It*

[40] Art. 4 ePD.
[41] "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee of the Regions "Building a European Data Economy" (COM(2017) 9 final), available at < https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2017:9:FIN>, accessed 20/03/2020.
[42] *Ibid.*

involves the generation, collection, storage, processing, distribution, analysis, elaboration, delivery, and exploitation of data enabled by digital technologies'.[43]

The Communication lists the four barriers to data mobility within the EU market:

- Data localisation restrictions put in place by Member States' public authorities that do not allow specific data to leave the country;
- IT systems vendors' obstacles (competition context);
- Fragmented EU legal patchwork in the area;
- Lack of trust due to security risks;
- Lack of cross-border data availability.

In the last years, the EU legislator has started developing legislative initiatives to eliminate such barriers. The Regulation on free-flow of non-personal data aims to remove data localisation restriction while the Privacy and Data Protection framework deals with risks originated while processing personal data.[44]

3.2.2.- Free Flow of Non-Personal Data Regulation

The Regulation on a framework for the free flow of non-personal data as adopted by the EU applies to the processing of electronic data other than personal data in the Union, which is (a) provided as a service to users residing or having an establishment in the Union, regardless of whether the service provider is established or not in the Union or (b) carried out by a natural or legal person residing or having an establishment in the Union for its own needs. In the light of the NPD Regulation, data shall be understood as data other than personal data as referred to in the GDPR.[45] As a result, its material scope of application encompasses any information that does not (including no longer) relate to an identified or identifiable natural person (see supra on the notion of personal data), i.e. non-personal data by nature or properly anonymised datasets. In the case of a data set composed of both personal and non-personal data, the NPD Regulation applies to non-personal data part of the data set. Contrary, where personal and non-personal data of a data set are inextricably linked, the NPD Regulation shall not prejudice the application of the GDPR.

General principles

In order to remove the barriers to the free movement of non-personal data within the EU, the NPD Regulation prohibits data localisation requirements, protects the availability of data for competent authorities and encourages the development of self-regulation facilitating the porting of data from one provider to another.

Firstly, the NPD Regulation contributes to the free movement of data within the EU and prohibits data localisation requirements, only with a very limited possibility of exceptions. National data localisation requirements, stemming from National legislation rules (mainly administrative guidelines or practices), dictate or influence the localisation within the

[43] *Ibid.*
[44] *Ibid.*
[45] Art. 3(1) NPD Regulation.

National territory of a Member states.[46] Such requirements mainly concern accounting documents, invoices, books and records, commercial letters, judicial records, national registries and archives and – broadly speaking – the servers hosting these data.[47] The NPD Regulation draws the attention that data localisation requirements represent a clear barrier to the free provision of data processing activities across the EU and to the internal market. As such, they shall be prohibited unless justified on grounds of public security in compliance with the principle of proportionality.[48]

In relation to this general prohibition of data localisation requirements, the Member States shall ensure that any existing data localisation requirement that is not compliant with the above-mentioned prohibition is repealed by 30 May 2021.[49]

Secondly, the NPD Regulation seeks to facilitate cross-border access to non-personal data by public authorities. The NPD Regulation stipulates that it does not affect the powers of competent authorities to request or obtain access to data in accordance with EU or national law, and that competent authorities cannot be refused access to data on the basis that the data are processed in another Member State.[50] In case after requesting access to a user's data, a competent authority does not obtain access and if no specific cooperation mechanism exists under EU law or international agreements to exchange data between competent authorities of different Member States, that competent authority may request assistance from a competent authority in another Member State.[51]

The notion of competent authority is a broad one and covers any authority or any other entity authorised by national law to perform a public function or to exercise official authority, that has the power to obtain access to data processed by a natural or legal person for the performance of its official duties, as provided for by Union or national law.[52]

Art. 7 of the NPD Regulation provides the requirements that shall be fulfilled in order to ensure cooperation between authorities.

Thirdly, the NPD Regulation seeks to contribute to efficiency of switching between service providers and to facilitate data portability. It is assumed that self-regulatory codes of conduct (hereinafter codes of conduct) shall be beneficial for these purposes[53], hence the NPD Regulation provides that the European Commission shall encourage service providers to complete the development of such codes by 29 November 2019 and to effectively

[46] For the definition, see: Commission staff working document on the free flow of personal data and emerging issues of the European data economy accompanying the document Communication Building a European Data Economy (COM(2017)9final, 5 <http://ec.europa.eu/newsroom/dae/document.cfm?doc_id= 41247> accessed 20/03/2020.
[47] During a study commissioned by the European Commission, more than 60 restrictions have been identified across 25 States. For a thorough analysis of data localisation requirements among Member States, see: Time.Lex, Spark, Tech4i2, 'Cross-border data flow in the Digital Single Market: data location restrictions' <https://ec.europa.eu/digital-single-market/en/news/cross-border-data-flow-digital-single-market-data-location-restrictions> accessed 20/03/2020.
[48] Art. 4(1) and Rec. 18 NPD Regulation.
[49] Art. 4(3) NPD Regulation.
[50] Art. 5(1) NPD Regulation.
[51] Art. 5(1) NPD Regulation.
[52] Art. 3(6) NPD Regulation.
[53] Rec. 29-30 NPD Regulation.

implement them by 29 May 2020.[54] Service providers shall be understood as natural and legal persons who provide data processing services.[55]

It is expected that codes of conduct will cover at least the key aspects that are important during the process of porting data, such as (i) the processes used for, and the location of, data back-ups; (ii) the available data formats and supports; (iii) the required IT configuration and minimum network bandwidth; (iv) the time required prior to initiating the porting process and the time during which the data will remain available for porting; (v) and the guarantees for accessing data in the case of the bankruptcy of the service provider. It is also expected that such codes will make clear that vendor lock-in is not an acceptable business practice, will provide for trust-increasing technologies, and will be regularly updated in order to keep pace with technological developments.[56] The Commission shall ensure that the codes of conduct are developed in close cooperation with all relevant stakeholders, including associations of small and medium-sized enterprises and start-ups, users and cloud service providers.[57]Any EU Member State has already implemented the EECC.

3.2.3.- Platform-to-Business Regulation

The Regulation on promoting fairness and transparency for business users of online intermediation services addresses legal challenges which originate from platform business model and represents one building block of the EC's legislative strategy for achieving the EU Digital Single Market. The P2BR foresees a list of measures that aim to enhance transparency and fairness, tempering the natural asymmetries that characterise the relationship between platform suppliers and vendors that use platforms to deliver their business or services. The P2BR focuses only on a specific type of platforms, namely, those offering services or products in competition with the platform business clients (i.e. Amazon).

The P2BR does not foresee a clear threshold, applying indistinctively to all types of platforms that fall in the criteria listed in Art. 2 P2BR.[58] According to Art. 2(2) P2BR intermediation services (platforms) that fall into the scope of application of this Regulation if:

1. *'(a) they constitute information society services within the meaning of the European Electronic Communication Code;*
2. *(b) they allow business users to offer goods or services to consumers, to facilitate the initiating of direct transactions between those business users and consumers, irrespective of where those transactions are ultimately concluded;*
3. *(c) they are provided to business users based on contractual relationships between, on the one hand, the provider of those services and, on the other hand, both those business users and the consumers to which those business users offer goods or services'*.[59]

[54] Art. 6(1), Art. 6(3) NPD Regulation.
[55] Art. 3(4) NPD Regulation.
[56] Rec. 29-31 NPD Regulation.
[57] Art. 6(2) of the NPD Regulation.
[58] Art.2(2) Regulation on promoting fairness and transparency for business users of online intermediation services.
[59] *Ibid*.

In the PIMCity deployment stage the services offer by the foreseen platform might fall under the definition of platform provided by Art.2 P2BR.

The P2BR follows two main principles, namely, **transparency** and **fairness**.
To enhance transparency in the platform to business context, providers of intermediation services are obliged to inform, through clear, unambiguous and readily available contractual terms and conditions (Art.3), about the treatment, the criteria used to rank their products, and the requirements to restrict, suspend or terminate their services (Art.4). When it comes to terms and condition, the P2BR specifies that ‚*providers of online intermediation services shall include a description of the technical and contractual access, or absence thereof, of business users to any personal data or other data, or both, which business users or consumers provide for the use of the online intermediation services concerned or which are generated through the provision of those services'* (Art.9).[60]

When it comes to the fairness principle, the P2BR establishes out-of-court dispute settlement mechanisms, such as internal complaint-handling system[61] (Art.9) and mediation (Art.12).[62] Contractual terms and conditions prepared by the intermediaries, thus, have to include a list of independent mediators that can be approached to settle disputes.

3.3.- Cybersecurity

3.3.1.- Directive on Security of Network and Information Systems (NIS)

The Directive on the Security of Network and Information Systems is the first EU legislative initiative in the cybersecurity area. It was developed as part of the EU Cybersecurity Strategy, which includes a set of binding and non-binding legal measures aimed at establishing a high standard of security across all EU Member states.

The NIS establishes standard procedures regarding the cooperation in the handling of security incidents both at national and EU level for private and public entities. It is characterised by a minimum level of harmonisation, leaving the possibility to the Member States to adopt or maintain higher standards for securing their network and information systems. Unfortunately, the different approach taken by the Member States in the implementation process of the NIS has led to legislative fragmentation, detrimental especially for those entities that operate cross-border within the EU.

Concerning private actors, the NIS applies to two different entities, namely, operators of essential services, and digital Services Providers. Art. 4 NIS defines operators of essential services as any private or public entity that falls under one of the categories referred to in Annex II of the NIS. These operators are considered necessary for the maintenance of critical societal and economic activities.

[60] Art.9(1) P2BR.
[61] Art. 11 P2BR.
[62] Art. 12 P2BR.

Contrary, digital service providers are legal persons providing one of the services listed in Annex III of the NIS such as online market place, online search engine or cloud computing service.[63] The necessity to ensure security protocols is given by the growing weight of their business within the EU economy and their cross-border nature.

The NIS specifies that if another EU legislation has already foreseen security measures, their provisions should prevail if they are proved to ensure at least equivalent security standards. Regrettably, what should be considered as equivalent standard is still uncertain.[64] Notwithstanding NIS provisions, many Member States have listed electronic communication providers as operators of essential service providers. Therefore, a security incident will require some operators to follow the procedures foreseen in the Code. In contrast, others will have to follow the protocols established in the NIS Directive. As a result, the different approaches taken by the Member States in the implementation of the NIS create legal fragmentation which results from harming those entities that operate cross-border.
Telefonica and **Fastweb**, crucial PIMCity's partners fall into the definition of Art.2 EECC, therefore, it is useful in the context of such deliverable to highlight the potential regulatory challenges that these two operators might have to face when providing their service cross-border.
Concurrently, Member States have also to put in place adequate measures for handling incidents. In particular, they have to adopt a national strategy on the security of network and information systems defining the strategic objectives and appropriate policy and regulatory measures to achieve a high level of security.[65]

3.3.2.- Cybersecurity Act

The EU Cybersecurity Act follows two main guidelines: the implementation of mandate and scope of the EU Agency ENISA, and the creation of an EU cybersecurity certification scheme for ICT products, ICT services, and ICT processes. According to the EC strategy, the adoption of EU standards in the cybersecurity context will enhance trust in the ICT products, services, and processes. As a result, the Cybersecurity Act will promote the reliability of EU companies within and outside the EU borders.

Art. 47 of the Cybersecurity Act states the application of EU cybersecurity certificates by industries and companies will be initially on a voluntary base. Each cybersecurity certification scheme will foresee different assurance levels that are considered the basis for users' confidence. The different assurance levels will depend 'on the risk associated with

[63] The definition included in Directive 1535/2015 refers to: service' . With service we intend any Information Society service, that is to say, any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services.
[64] Rec. 9 NIS: *'Certain sectors of the economy are already regulated or may be regulated in the future by sector-specific Union legal acts that include rules related to the security of network and information systems. Whenever those Union legal acts contain provisions imposing requirements concerning the security of network and information systems or notifications of incidents, those provisions should apply if they contain requirements which are at least equivalent in effect to the obligations contained in this Directive.'*
[65] Art. 8 NIS.

the intended use of the ICT product, ICT service or ICT process, in terms of the probability and impact of an incident.'

The Cybersecurity Act and in particular, the development of certification schemes might have an impact on PIMCity and its partners. The first schemes developed are going to be in the 5G network and cloud computing services area. Therefore, the adoption of cybersecurity certification schemes for ICT products, services, and the process might prove useful for PIMCity partners.

3.4.- Telecom

3.4.1.- European Electronic Communications Code

The Directive establishing the European Electronic Communication Code has been adopted on the 11th of December 2018 and needs to be implemented by the Member States in two years (December 2020). The EECC amends four different Directives[66] and governs all aspects involving providers of electronic communication networks and their competent national authorities.

In the PIMCity project, it is useful to analyse two sets of EECC's provisions: consumer protection and security ones.

From a consumer protection perspective, the EECC intends to increase consumer choice, providing high quality and innovative services at a lower price, and to enhance security and confidentiality of communications. To achieve such an ambitious purpose, the EECC has broadened its scope of application, aiming to establish an equal playing field between "Traditional" and "New" operators (OTTs). Differently from previous legislation Art. 2 EECC includes new and traditional operators which fall in the EECC scope as providers of electronic interpersonal communication services (ICS).[67] The traditional operators are defined as 'Number-Based ICS ('NB ICS') providers',[68] while OTTs are listed as 'number-independent ICS ('NI ICS')'[69] providers' (e.g. WhatsApp).

Notwithstanding the inclusion of new operators in the scope of the EECC, from a consumer protection angle, the initial aim to establish the same playfield between competing ICS providers is not fully achieved. For example, when it comes to contract duration and termination, Art. 105 EECC excludes NB ICS providers from the consumer protection provisions.[70]

[66] Specifically, the Code amends Directive 2002/19/EC, 2002/20/EC, 2002/21/EC and 2002/22/EC.
[67] Art. 2(5) EECC: *'interpersonal communications service' means a service normally provided for remuneration that enables direct interpersonal and interactive exchange of information via electronic communications networks between a finite number of persons, whereby the persons initiating or participating in the communication determine its recipient(s) and does not include services which enable interpersonal and interactive communication merely as a minor ancillary feature that is intrinsically linked to another service;'.*
[68] Art. 2(6) EECC.
[69] Art. 2(7) EECC.
[70] Art. 105 EECC.

When it comes to security, Arts. 40 – 41 EECC are dedicated to the provisions governing the measures that electronic communication providers and Member States should put in place to secure their networks. On the one hand, providers of public electronic communication networks have to inform their users about any potential security threat that might affect their service, and of the measures taken to ensure the security of communications.[71] On the other hand, Art. 41 EEC requires in case of an incident, mandatory collaboration between the Member States and National Center of Security Incident Recovery Teams (CSIRT) and other relevant national authorities.[72]

3.5.- Consumer Protection

3.5.1.- Digital Content and Digital Services Directive

The Directive on contracts for the supply of Digital Content and Services regulates contracts for the supply of digital content and services that can be concluded not only in exchange of price but also in exchange of the consumers' personal data. The DCD does not pay attention to digital content ownership but instead on the remedies that consumers have when contracting for the supply of digital content using their personal data as a form of payment.

Rec. 19 and Art. 2 DCD provide the definition and concrete examples of what should be considered as digital content and service. Digital content should be regarded as data that are produced and supplied in digital form. At the same time, services should be intended as those that allow a consumer to create, process, store, access, share or allow interaction with data. This definition includes, *inter alia, computer programmes, applications, video files, audio files, music files, digital games, e-books or other e-publications, and also digital services,* which allow the creation of, processing of*, accessing* or storage of data in digital form, *including software-as-a-service, such as video and audio sharing and other file hosting, word processing or games offered in the cloud computing environment and social media*.'[73]

The DCD, dealing with personal data makes explicit reference to GDPR. In particular, Rec. 38 DCD states that '*any processing of personal data in connection with a contract falling within the scope of this Directive is lawful only if it conforms with the provisions of Regulation (EU) 2016/679 relating to the legal grounds for the processing of personal data'.*[74] Concerning personal data, it is useful to differentiate between personal data provided as a form of remuneration for the conclusion of a contract, and data *presented to the trader for supplying the digital content or digital service or for allowing the trader to comply with legal requirements to which the trader is subject, and the trader does not process that data for any other purpose'.*[75] When dealing with personal data *any supplier of digital content or service*, has to comply with the GDPR requirements.[76] Moreover, the DCD clarifies that a

[71] Art.40 EECC.
[72] Art.41 EECC.
[73] Rec. 19 DCD.
[74] Rec. 38 DCD.
[75] Art. 3 DCD.
[76] Rec. 69 DCD.

lack of compliance with the GDPR's requirements affects the conformity requirements of the contract falling into the scope of DCD.[77]

The definition provided in the DCD of digital content, data which are produced and supplied in digital form; and also the one of digital service, might require PIMCity partners to take into account the provisions in case they will decide to offer PIMCity functionalities to consumers.

3.6.- Copyrighted data

At the EU level, copyright issues are mainly addressed by the InfoSoc Directive which aims at harmonising legislation across Member States and at transposing the international acquis into the EU legislative order[78]. This Directive provides basic rules with regards to copyright in the EU. Besides, the EU adopted the CDSM Directive which provides for some relevant exceptions to copyright with regards to scientific research, text and data mining, teaching activities and preservation of cultural heritage. The CDSM Directive came into force on 7th June 2019 and shall be implemented by Member States by 7th June 2021. It shall be noted however that despite the existence of the overarching legislative texts, there is currently no common, fully harmonised legal framework for copyright within the EU.[79] Instead, the matter remains largely influenced by national legislation, international agreements and the case-law of the CJEU.

3.6.1.- Scope of application

Copyright grants the author of a **work,** which is **original,** exclusive rights over its **embodiments**. *First*, copyright covers literary and artistic works, a notion that encompasses *every production in the literary, scientific and artistic domain, whatever may be the mode or form of its expression, such as books, pamphlets and other writings.*[80] The scope of this

[77] Rec. 48 DCD: '…*One example could be where a trader explicitly assumes an obligation in the contract, or the contract can be interpreted in that way, which is also linked to the trader's obligations under Regulation (EU) 2016/679. In that case, such a contractual commitment can become part of the subjective requirements for conformity'.*

[78] The main international agreements on copyright are: the Berne Convention for the Protection of Literacy and Artistic Works of 9 September 1886 (hereinafter as the Berne Convention), the Universal Copyright Convention adopted in Geneva on 6 September 1952 and revised in Paris on 26 July 1971, the Agreement on Trade-Related Aspects of Intellectual Property Rights adopted in Marrakech on 15 April 1994 (Annex 1C of the Agreement establishing the World Trade Organisation) and the World Intellectual Property Organisation Copyright Treaty adopted in Geneva on 20 December 1996. Directive 2001/29, on the other hand, echoes the main points of these major international agreements.

[79] For more information on the legislative framework surrounding copyright, see: Benoit Van Asbroeck, Julien Debussche and Jasmien César, 'Bird & Bird White Paper - Building the European Data Economy - Data Ownership' 61 <https://www.twobirds.com/en/news/Art.s/2017/global/ data-ownership-in-the-context-of-the-european-data-economy> accessed 20 March 2020.

[80] Art. 2 of Berne Convention also refers to *lectures, addresses, sermons, and other works of the same nature; dramatic or dramatico-musical works; choreographic works and entertainments in dumb show; musical compositions with or without words; cinematographic works to which are assimilated works expressed by a process analogous to cinematography; works of drawing, painting, architecture, sculpture, engraving and lithography; photographic works to which are assimilated works expressed by a process analogous to photography; works of applied art; illustrations, maps,*

notion is sufficiently broad to include data, as they are generated and/or collected in a written format. *Second*, the work at stake must be original, i.e. reflect the author's personality, where he or she has been able to express his or her creativity by making free choices.[81] Originality also implies an intellectual effort from the author. Certain data are not likely to be considered original as they do not meet the latter criteria. In case certain data reflect their author's personality, they might be protected by copyright. *Third*, copyright only protects the embodiment of the work. In other words, it is crucial to distinguish the tangible expression of the work at stake from the underlying information. Only the former can be protected under copyright law while the factual data remain freely usable by anyone.[82]

3.6.2.- General principles

Copyright grants the right holder the **exclusive prerogatives** to (1) reproduce, (2) communicate to the public and (3) distribute the protected work.[83] Before performing any of these acts, third parties should, therefore, seek authorisation from the author, or rely on one of the various exceptions listed in the InfoSoc Directive or the CDSM Directive. It should also be noted that, while copyright is initially granted to the author of the work – i.e. the natural person who actually expressed its creativity – it may subsequently be transferred or licensed. This may not only be done in bulk, but also by fractioning the components of the copyright among different transferees/licensees, by limiting the territorial scope of the transfer/license to definite territories or by stipulating specific exploitation modalities for each component.

3.6.3.- Exclusive rights of the author

First, the author has the exclusive right to *authorise or prohibit direct or indirect, temporary or permanent **reproduction** by any means and in any form, in whole or in part.*[84] As often emphasised by the case-law of the CJEU, the notion of reproduction must be interpreted as to encompass the mere reproduction for technical purposes (*e.g.* cache copies, conversion in a different format, back-up copies preventing data loss, screen buffer, etc.)[85] If the PIMCity partners are to process copyrighted data, it is, therefore, reasonable to assume that there will be a reproduction within the meaning of the InfoSoc Directive. Such reproduction will have to be authorised by the author or the relevant right holder or will have to be relied on one of the exceptions foreseen by the InfoSoc Directive or the CDSM Directive (see *infra*). Second, the author also has the exclusive right to *authorise or prohibit any **communication to the public** of his/her work, by wire or wireless means, including the making available to the public in such a way that members of the public may access them*

plans, sketches and three-dimensional works relative to geography, topography, architecture or science. It is essential to underline that this list is no limitative.

[81] On the notion of originality as outlined above, see *a.o.*: *Infopaq v. Danske Dagblades Forening*, Case C-5/08, para. 32-39; *BSA v. Ministerstvo Kultury*, Case C-393/09, para. 45; *Premier leage*, Case C-403/08 and C-429/08, para 97 and 155; *Painer v. Standard*, Case C-145/10, para. 90-92; *Football Dataco*, Case C-604/10, para. 37; *Nintendo v. PC Box*, Case C-355/12, para. 21.

[82] This ties back to the theory according to which copyright does not cover abstract ideas, even if they are original. Instead, copyright only deals with ideas that have been materialised in a concrete form and, therefore, with the way the authors have concretely expressed this idea.

[83] Arts. 2, 3, and 4 Directive 2001/29, respectively.

[84] Art. 2 of the InfoSoc Directive.

[85] See, for instance: CJEU, *Infopaq International v. Danske Dagblades Forening*, case C-5/08, para 51, CJEU, *Premier League*, case C-403/08, para. 159.

from a place and a time individually chose by them.[86] This is less likely to fall within the categories of acts performed by the PIMCity partners. Third, the author has the exclusive right to *authorise or prohibit any form of* ***distribution*** *to the public by sale or otherwise.*[87] Again, this is less likely to fall within the categories of acts performed by the PIMCity partners.

3.6.4.- Relevant exceptions to the rights of the author

The InfoSoc Directive and the CDSM Directive stipulate several potentially relevant exceptions, i.e. cases in which authorisation of the right holder may be not necessary to perform the relevant actions.

First, the InfoSoc Directive introduces a mandatory exception for *temporary acts of reproduction which are transient or incidental and an integral and essential part of a technological process whose sole purpose is to enable (1) a transmission in a network between third parties by an intermediary or (b) a lawful use.*[88] This exception only applies when the temporary acts of reproduction have no independent economic significance. It mainly covers caching and browsing operations which, otherwise, would be conditional upon the authorisation of the author of the work which is cached or displayed.[89]

Second, the InfoSoc Directive also gives Member States the possibility to introduce other exceptions to the exclusive reproduction right of the author. It offers the possibility to limit or restrict this right in case the copyrighted work is used *for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved.*[90] However, it shall be recalled that not every Member State has transposed this optional provision in its national legislation, and, for those that did, significant divergences exist.[91] Additionally, the requirement that the copyrighted work must be used solely for scientific research might prevent from benefiting from this exemption. Similarly, the non-commercial criteria might also raise problematic issues. The particular applicability will need to be assessed on a case-by-case basis, taking into account the work at stake and the scope of the exemption as implemented in national legislation interpreted by domestic courts.

Since the adoption of the CDSM Directive it is relevant that certain provisions of the InfoSoc Directive shall be interpreted **without prejudice to the exceptions and limitations newly provided by the CDSM Directive.**[92] Hence it is necessary to clarify the regime introduced

[86] Art. 3 InfoSoc Directive.
[87] Art. 4 InfoSoc Directive.
[88] Art. 5(1) InfoSoc Directive.
[89] Rec. 33 Directive 2001/29 clarifies that *this exception should include acts which enable browsing as well as acts of caching to take place, including those which enable transmission systems to function efficiently, provided that the intermediary does not modify the information and does not interfere with the lawful use of technology, widely recognised and used by industry, to obtain data on the use of the information*.
[90] Art. 5(3)a InfoSoc Directive.
[91] See on that point: Jean-Paul Triaille and others, *Study on the Legal Framework of Text and Data Mining (TDM).* (Publications Office 2014) 368–370 <http://bookshop.europa.eu/uri?target=EUB:NOTICE:KM0313426:EN:HTML> accessed 20/03/2020.
[92] Art. 24(2)b CDSM Directive.

by the CDSM Directive. Besides, the latter document establishes some other potentially relevant exceptions. Both issues are further addressed below.

Firstly, it is clarified by the CDSM Directive that Member States shall provide for an exception to the exclusive right to authorise reproduction as revealed above[93] for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.[94] In this context it is relevant that research organisation shall be understood as university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research: (i) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or (ii) pursuant to a public interest mission recognised by a Member State; in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence upon such organisation.[95] Conversely, organisations upon which commercial undertakings have a decisive influence allowing such undertakings to exercise control because of structural situations, such as through their quality of shareholder or member, which could result in preferential access to the results of the research, should not be considered research organisations for the purposes of this Directive.[96] Research organisations shall benefit from such an exception also when their research activities are carried out in the framework of public-private partnerships. It is explicitly provided that research organisations shall be able to rely on their private partners for carrying out text and data mining, including by using their technological tools.[97]

In addition, Member States shall provide for an exception or limitation to the exclusive right to authorise reproduction in order to allow the digital use of works and other subject matter for the sole purpose of illustration for teaching, to the extent justified by the non-commercial purpose to be achieved, on condition that such use: (i) takes place under the responsibility of an educational establishment, on its premises or at other venues, or through a secure electronic environment accessible only by the educational establishment's pupils or students and teaching staff; and (ii) is accompanied by the indication of the source, including the author's name, unless this turns out to be impossible.[98] However, Member States may provide for some exceptions to this exception or limitation.[99]

The exception or limitation should benefit all educational establishments and emphasise that it should apply only to the extent that the uses are justified by the non-commercial purpose of the teaching activity. While deciding whether the activity is non-commercial, the organisational structure and the means of funding should not be the decisive factors.[100] The

[93] Art. 2 InfoSoc Directive.
[94] Art. 3(1) CDSM Directive.
[95] Art. 2(1) CDSM Directive.
[96] Rec. 12 CDSM Directive.
[97] Rec. 11 CDSM Directive.
[98] Art. 5(1) CDSM Directive.
[99] Art. 5(2) CDSM Directive.
[100] Rec. 20 DCSM Directive.

use of works or other subject matter under the exception or limitation should be limited to what is necessary for the purpose of such activities.[101]

Besides, Member States shall provide for an exception or limitation to the exclusive right to authorise reproduction for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.[102] The latter exception or limitation shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their right holders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.[103]

It remains to be seen whether the particular subjects will be able to benefit from these exceptions, depending on their status, activities, access to work at stake, transposition in the Member States, including but not limited. All of the circumstances will need to be assessed in the light of the criteria on a case-by-case basis.

3.7.- Ethical Guidelines

Ethics principles are increasingly becoming a fundamental tool to interpret and adapt legal principles to the fast modifications that are increasingly shaping our society. Legal and ethical principles have been influencing each other for many years. On the one hand, laws offer the legislative setting upon which individuals, society and authorities carry out their activities. On the other hand, ethics provides interpretative tools to guide policymakers to build the normative architecture. To sum up*, 'Ethical considerations can play a part in system governance by shaping the actions of people, imposing constraints and providing guidelines for the development and design of technology'*.[104]

In the PIMCity context, the ethical guidelines that will be described in this deliverable and applied during the whole project aim to guide each partner involved in the PIMCity activities.

3.7.1.- Fundamental Moral Principles

The four pillars of ethics upon which the legislative architecture of our society is built are autonomy, justice, beneficence and non-maleficence. Upon these four principles, an additional secondary principle has been created: the principle of responsibility. According to the different context where they are applied, some principles result to be more relevant than others. Nonetheless, each one of such principle should always be taken into account when developing new technological solutions.[105]

1. **The principle of autonomy.** Every individual has the fundamental right to self-determination. As a result, each one of us has positive and negative obligations.
   a. *Negative obligation*: our action should not be detrimental for other individuals.

[101] Rec. 22 DCSM Directive.
[102] Art. 4(1) CDSM Directive.
[103] Art. 4(2) CDSM Directive.
[104] Verhenneman G, Vedder A, WITDOM, D6.1 – Legal and Ethical framework and privacy and security principles, p. 39, 2015, available at: http://www.witdom.eu/deliverables, accessed 20/03/2020.
[105] *Ibid.*

b. *Positive obligation*: each individual should rely on a respectful treatment when revealing information and taking independent decisions.
The fundamental right to privacy and data protection and the obligations developed in different legislation to preserve such right are strictly linked to this ethical principle.

2. **The principle of justice**. The principle of justice entails that all individuals 'are entitled to have the same degree of attention and moral concern.' As a result, it requires that each individual has to be treated with fairness taking into account their different needs, contributions, and vulnerabilities. The ethical principle of justice is embedded in the privacy and data protection framework. The principle of fairness should apply to avoid power asymmetries between data subjects and controllers lead to potential abuses.
3. **The principle of beneficence**. The principle of beneficence requires that all individuals have to contribute to personal and societal well-being. In the privacy and data protection context, development of such an ethical principle is embedded, for example, in those provisions requiring entities in charge of processing activities to ensure the security of the data subjects.
4. **The principle of non-maleficence.** This principle of non-maleficence is originated from Hippocrates' oath *primum non nocere* (first do not harm). Differently from the principles mentioned above, this principle has been mainly developed in the biomedical context. Nonetheless, since the principle of non-maleficence requires individuals with a duty do not have to cause harm to other vital health and security interests, we can apply it indirectly to the controller data subject relationship.
5. **The (secondary) principle of responsibility.** The principle of responsibility requires that each individual involved in a given project should behave and fulfil its moral obligations. Such duties stem from its role in the project, at the best of his/her abilities. In any given project, the principle of responsibility implies that each party involved have the responsibilities to carry out its tasks, and the consequences originated from them.[106]

The five moral principles listed above will be taken into account, using them as an ethical baseline during the whole PIMCity project's activities.

3.7.2.- EDPS' Ethics Advisory Group 2018 Report, *Towards a digital ethics*

From an ethical perspective, one of the significant challenge of the PIMCity project concern the processing activities involving personal data. Therefore, it is useful to highlight the latest ethical developments in the privacy and data protection area.

Established following the Regulation on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on

[106] Verhenneman G, Vedder A, WITDOM, D6.1 – Legal and Ethical framework and privacy and security principles, p. 39, 2015, available at: http://www.witdom.eu/deliverables, accessed 20/03/2020.

the free movement of such data[107] and on the basis of Art. 16(2) TFEU,[108] the European Data Protection Supervisor (EDPS) is an independent EU authority to ensure that the fundamental rights and freedoms of individuals – in particular their privacy – are respected by the EU institutions and bodies. Among other crucial responsibilities, the EDPS provides opinions and advice to EU institutions and agencies regarding the impact that upcoming legislative initiatives might have on individuals right, such as the right to privacy and data protection.

Before the GDPR approval, the EDPS published a Report describing how the effects the new technologies are reshaping the relationship between technology and human values.[109] The Report, prepared by the Ethics Advisory Group (EAG), follows the EDPS 2015 Opinion Toward a new Digital Ethics.[110] The EDPS' Report also provides guidelines to policymakers on how to combine ethical principles with challenges coming from digital innovation.

The five political (non-binding) recommendations to support and develop the European values based on the ones embedded in the data protection framework are:

1. The essential and inviolable human dignity has to be preserved, regardless of any change that might occur in society
2. Personhood, with his or her moral values and social and cultural characteristics, cannot be taken apart from his or her personal data;
3. Freedom of choice has to remain a pillar of society. Therefore, autonomous decision making cannot undermine such a principle;
4. In the context of profiling, accountability should be fostered to avoid any form of discrimination;
5. Data commoditisation can lead to potential tension if human moral values are not taken into account.[111]

Taking into account the framework described in this deliverable D1.1, KUL will develop a concrete list of legal requirements for the PIMCity platform and its components (deliverable D7.2). The final version of the ethical and legal requirements will be provided in the deliverable D7.5, due date month 30.

[107] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC, PE/31/2018/REV/1, OJ L 295, 21.11.2018.

[108] Consolidated version of the Treaty on the Functioning of the European Union, OJ C 326, 26.10.2012, p. 47–390; Art. 16(2) TFEU: *'The European Parliament and the Council, acting in accordance with the ordinary legislative procedure, shall lay down the rules relating to the protection of individuals with regard to the processing of personal data by Union institutions, bodies, offices and agencies, and by the Member States when carrying out activities which fall within the scope of Union law, and the rules relating to the free movement of such data. Compliance with these rules shall be subject to the control of independent authorities.'*

[109] Ethics Advisory Group 2018 Report, Towards a digital ethics, available at < https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf>, accessed 20/03/2020.

[110] EDPS, Opinion 4/2015, Towards a new digital ethics Data, dignity and technology, 2015, available < https://edps.europa.eu/sites/edp/files/publication/15-09-11_data_ethics_en.pdf> accessed 20/03/2020.

[111] Ethics Advisory Group 2018 Report, Towards a digital ethics. p.20

4.- PIMS requirements for the PDK

The PDK is a collection of generic, reusable and extensible components that are fundamental for building new PIMS or update existing ones without frictions, built on privacy and secure-by-design principles.

In an ever-evolving industry, it is crucial that all PDK components are defined, designed, developed, tested and released using an Agile Development methodology (such as Scrum) because such framework will allow the team to have adaptive planning, evolutionary development, early delivery, and continual improvement, which will also encourage flexible responses to change.[112]

The following requirements have been identified and agreed by entities in the Consortium running commercial PIMS, ensuring that they are in line with what a commercial entity would demand in the market.

4.1.- PIMS Functionality

PIMS, also referred to as personal data stores, personal data spaces, or personal data vaults, are systems that empower individuals to take control of their personal data by enabling them to gather, store, update, and manage contracts to share them with third parties. In fact, one key functionality of PIMS is that they let individuals manage their consent to third-parties for access to their personal data, so they can allow, deny or withdraw their consent to third parties accessing all or part of their personal information, and set the terms and conditions for such sharing to take place, e.g. share location to receive location-sensitive advertisements if and only if the user is paid a certain monthly price. Consequently, PIMS can facilitate compliance with existing privacy laws by making it easier for organizations to gain effective consent of users and manage contracts under which they have access to personal information.

One of the objectives of PIMCITY's PDK is to provide functionality that helps PIMS provide value to their users, be they individuals organizing and sharing their data, or entities interested in getting access to their personal information. Consequently, the PDK should explore the given functionalities:

- Gather and organize personal information from different sources, as well as a scalable framework to add new data sources
- Provide a scalable data taxonomy to organize and catalogue personal data
- Securely store personal data
- Manage the users' consent to share personal data,
- Provide marketplace functions to let users trade with their data
- Manage the terms and contracts signed with third parties for such data sharing to take place
- Manage provide transparency to data transactions and money transfers

[112] https://www.c-sharpcorner.com/article/why-agile-is-important-in-software-development/

- Provide and manage secure access to the users' shared data
- Track both the origin of data and data shared with third parties to increase the security of data exchanges
- Provide privacy preserving logs of activity and dashboard to the different types of users (data sellers and data buyers)

4.2.- Technical standards

In addition to the aforementioned functional requirements, to be able to deliver components that can be integrated in the PDK, each of them has also to meet standards on Documentation, Versioning, Commands, Tests and Coverage, Code Quality, and Monitoring. Such standards are specified below.

4.3.- Documentation

4.3.1.- License

The published component must have an open source license. Otherwise it is not open source.[113]

4.3.2.- Readme

Simple and welcoming explanation about why the project is useful and how to get started. This is one of the first things developers read and it is key to make it a good resource for them.[114]

4.3.3.- Contributing Guidelines

Documentation for people in the community that want to go a step forward and start contributing to the PIMCity in some way. Remember that code is not the only way to contribute to the project, take into account the following forms of contributing: designing UI solutions, writing down or improving documentation, adding examples and tutorials, helping with the organization of the project.[115]

4.3.4.- Code of conduct

A code of conduct is a document that establishes expectations for behavior for your project's participants. Adopting, and enforcing, a code of conduct can help create a positive social atmosphere for the PIMCity community.[116] The usual place for this document is the code repository, where it can be versioned and contributors can read it and commit to it.

In addition to communicating your expectations, a code of conduct describes the following:

- Where the code of conduct takes effect (only on issues and pull requests, or community activities like events?)

[113] https://medium.com/coinbundle/what-is-open-source-code-8579cf7a8bb0
[114] https://gist.github.com/PurpleBooth/109311bb0361f32d87a2
[115] https://github.blog/2012-09-17-contributing-guidelines/
[116] https://github.blog/2015-07-20-adopting-the-open-code-of-conduct/

- Whom the code of conduct applies to (community members and maintainers, but what about sponsors?)
- What happens if someone violates the code of conduct
- How someone can report violations

4.3.5.- Issue tracker

There has to be documentation regarding how to report issues, and a specific section for security concerns.[117]

4.3.6.- Pull requests

In order to submit a change to the code base, all contributors must create a Pull Request[118] answering the following questions:

- What does this Pull Request do?
- Are there points in the code the reviewer needs to double check?
- Why was this Pull Request needed?

After that, the contributor must include this checklist, making sure all points are met:

- "I have read the Contribution guidelines"
- "I have read the Code of Conduct"
- "I have updated the documentation accordingly"
- "I have added tests to cover my changes."

Finally, for the changes to be merged, the Pull Request must meet the next criteria:

- At least one team member reviewed and approved the changes.
- All automatic tests pass.
- Code Coverage is not less than the one that the project had before the changes.

4.4.- Versioning

Tools included in the PDK must follow the semantic versioning scheme: MAJOR.MINOR.PATCH. This is key to have a good level of evolvability.[119]

To be able to iterate fast, and having a not so stable API in the beginning, the best way to embrace semantic versioning is to start with version 0.1.0. All subsequent releases may update MINOR or PATCH values.

When the components reach a certain level of maturity and are tested long enough, the first production-ready release changes to 1.0.0. Since then, the following scheme has to be adopted:

[117] https://www.donedone.com/use-issue-tracking-software/
[118] https://blog.carbonfive.com/why-write-good-pull-requests/
[119] https://www2.le.ac.uk/services/research-data/old-2019-12-11/organise-data/version-control

Given a version number MAJOR.MINOR.PATCH, increment the:

MAJOR version when you make incompatible API changes,

MINOR version when you add functionality in a backwards compatible manner, and

PATCH version when you make backwards compatible bug fixes.

Additional labels for pre-release and build metadata are available as extensions to the MAJOR.MINOR.PATCH format.

4.5.- Development Tools

Every component in the PDK must have a set of development tools or shell commands that assist developers to run a local version of the service or program, test the core functionalities, and understand the current code coverage. Optionally, additional commands to automate setup, lint, fix code, run different types of tests (unit tests, integration tests, CI context) and build the service or program to a deployable state are all welcome.

We encourage developers to use Docker and Docker Compose whenever possible so that the process of setting up and running a service with dependencies is friction-less.

4.6.- Tests and coverage

Before merging any pull requests (or merge requests in GitLab), the test suite must be run, each change should not brake previous working functionality or tests, and code coverage must be above predefined and discussed coverage percentage. To automate this process, PIMS projects are encouraged to use a CI platform that can handle the build process, lint the code, run the test suite, calculate the coverage and fail (preventing merge) when any of the checks fail.

4.7.- Code quality

To build a generic, reusable, and extensible component, we need to spend time designing it, but without defining a code style to write clean-code previous efforts can turn useless.[120] For this matter, PIMS developers must decide in advance which code style the team must follow, the rules (with thresholds), and the tool that it is going to check them.

In addition to the code style, a static code analyser must check for the following metrics to provide readable code:

- ABC size: based on assignments, branches, and conditions. The value per method should not be higher than 15.
- Cyclomatic complexity: counts the decision points inside a piece of code. It should not be higher than 6 per method.
- Parameter lists: counts the number of parameters, including named parameters. Each method should have no more than five parameters.

Finally, PIMS components should follow the SOLID principles:

[120] https://medium.com/@cleverti/why-is-code-quality-such-a-big-deal-for-developers-91bdace85d44

- Single Responsibility Principle: every class or module should have responsibility for a single part of the functionality
- Open/Closed Principle: Classes, modules, and functions must be open for extension but closed for modification.
- Liskov Substitution Principle: Objects in a program should be replaceable with instances of their own subtypes without breaking the program.
- Interface Segregation Principle: States that no client should be forced to depend on methods it does not use
- Dependency Inversion Principle: High-level modules should not depend on low-level modules. Both should depend on abstractions. Abstractions should not depend on details. Details should depend on abstractions.

4.8.- Monitoring

Monitoring and alerting enables a system to tell us when it's broken, or perhaps to tell us what's about to break. When the system isn't able to automatically fix itself, we want a human to investigate the alert, determine if there's a real problem at hand, mitigate the problem, and determine the root cause of the problem.[121]

The PIM Systems are not going to implement a full-stack monitoring and alerting system. Still, they need to have basic instrumentation to allow developers and system integrators to achieve that goal. The following tools are needed:

4.8.1.- Logging

A set of primitives to log messages, one for each of Unix Security levels. By default, every message is logged to stdout.

4.8.2.- Telemetry

In the case of services that interact with other services or with the final user and to ensure a satisfactory experience, PIMS must instrument and keep track of Latency, Traffic, Errors and Saturation. By default, these metrics are going to be logged to stdout.

4.8.3.- Alerting

As developing systems and integrating them is done by humans, there is a possibility that a developer can introduce a bug in a change. PIMS components must be prepared to alert when this happens, so that system administrators can act proactively. Alerting also occurs when a threshold in Latency, Traffic or any other metric is not met. By default, alerting is logged to stdout.

[121] https://www.swicktech.com/SWICKtech/Resources/Blog/Why-do-my-computer-systems-need-monitoring.htm

5.- Requirements for a user-centric data economy

In addition to the legal and user requirements a series of additional requirements stem from the more general economic environment in which a PIMS, built upon the PDK, will have to operate. The aforementioned economic environment is one in which data is an essential production factor (just as capital, labour or land has been until now). In this context, a user-centric data economy (hereinafter, UCDE) will require that individuals are compensated by companies for their data in proportion to the benefits that such data produce for the overall economy.

But why do we need a user centric data economy? Paying people for their data seems to be at least one step ahead from discussions about data in our times, which go mostly in the direction of data protection and privacy. Still the arguments in favour of alternative data economies are mounting fast:

- Paying for data puts economic pressure on online services to apply *data minimisation principles*, i.e., to collect and process only the minimum amount of data necessary for their operation. Data minimisation is mentioned in GDPR and other data protection laws, but is seldom applied in practice. Indeed, since data is by now a major asset, companies collect all they can, far and beyond their actual information needs, in the hope that future versions of their service, or new business models, will allow them to eventually monetize the extra amount collected. This, in conjunction with the fact that that *collection and processing of data costs close to zero,* are fuelling greedy "all-you-can-eat" practices, which harm privacy and can eventually cause a privacy-trust related tragedy of the commons in the Internet ecosystem. By having to pay for data, "parasitic" services, such as trackers that compile lists of anything from suspected alcoholics and HIV positive individuals, to active police-officers (see for instance https://money.cnn.com/2013/12/18/pf/data-broker-lists/, last accessed January 2020), would go out of business, whereas valuable services such as search, maps, etc. would proceed uninhibited and even benefit, as explained later.

- Beyond its negative effects on privacy, the current economic model around data has led to market failures in the form of large data monopolies and oligopolies, and may even become a threat to employment in the future due to job loss from data-driven automation. Paying people for their data could, therefore, be an alternative to labour-based compensation in a future in which most work will be done by machines.
- Going a step further, paying for data should not be seen as harmful to business, for the simple reason that the online services market is certainly not a zero-sum market — increasing the profit of users does not have to necessarily harm the profits of online services. In fact, by providing compensations for data according to data quality and usefulness, online services can acquire more data, and of higher quality, than the data they intrusively collect today, and by doing so, increase their revenues and the utility for their users.

Bootstrapping a user-centric data economy will require coming up with fair rules for providing *explicit monetary benefits* for the providers of data, without whom, data-driven business models and their supporting machine learning algorithms, have next to zero value. Paying users for data has the potential to smooth out some/much of the existing tension around user privacy vs. utility for online services. Moreover, paying for data should not be

seen as harmful to business since this market is certainly not zero-sum. Actually, providing compensation is the way for getting more and higher quality data that can in turn produce additional revenue for online services and better utility for their users, especially when this compensation is at least partially attached to the utility that each piece of data has for the purchasing entity.

As a consequence, a UCDE imposes certain requirements to the PDK. As a matter of a fact, the following figure summarize the main challenges that can be preliminarily foreseen for such data economy vision to become a reality:

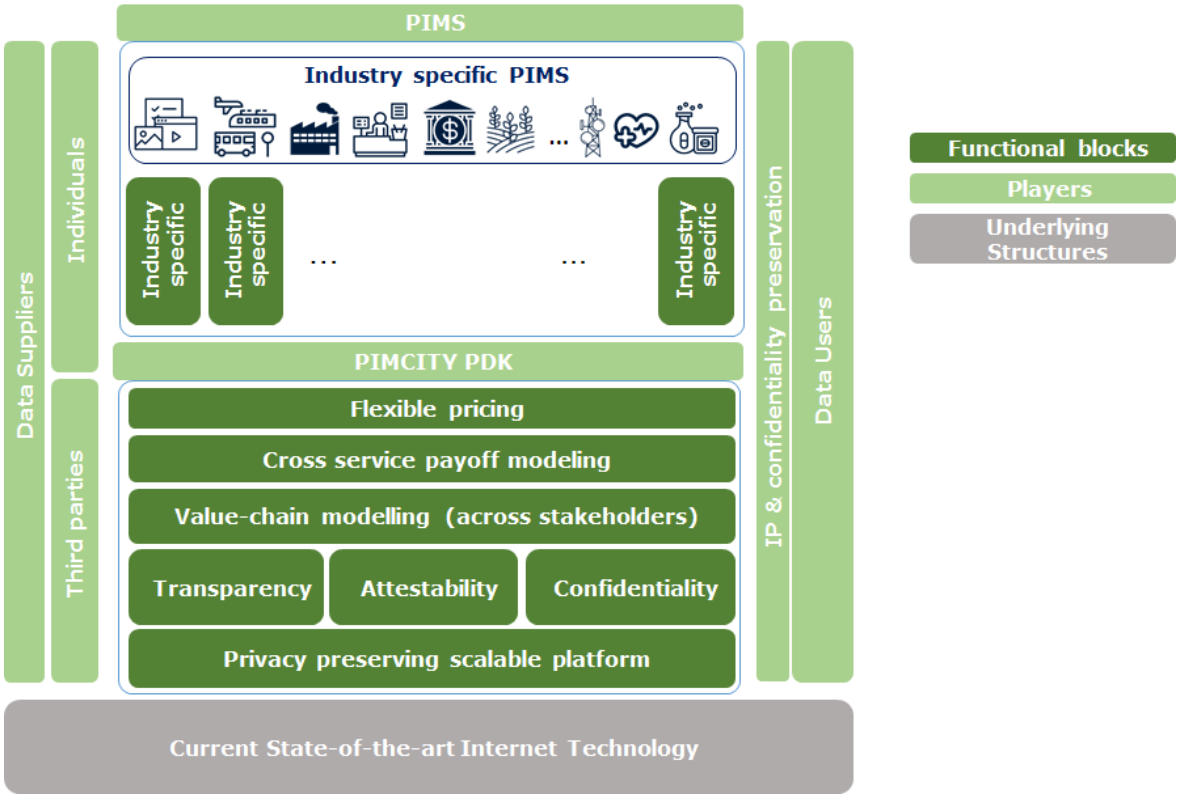

Figure 7: Requirements for a user centric data economy (UCDE)

Going from left to right, Figure 7 depicts the functional and business flow of a data transaction, the involved entities, and the functional blocks that the PIMS must provide, from a UCDE perspective. In a nutshell, a PIMS claiming to be user centric should:

- Be cross service compatible, i.e., provide a set of tools that are useful to different industries and verticals.
- Allow for flexible pricing models.
- Guarantee a fair payoff allocation among entity types (service, users, and third parties) involved in data transactions.
- Support a fair payoff allocation among different users, in proportion to the utility of their respective data to the final outcome of an algorithm.
- Be scalable in terms of volume of users and data transactions.
- Provide the different entities in the ecosystem with tools that can attest that monetary rewards for data are accurately and transparently computed.
- Respect the privacy, confidentiality and intellectual property of all involved entities.
- Be incrementally deployable over the existing technologies and services

The following sections elaborate on each of these functional blocks and present some examples in order to understand their implications for the PDK technical requirements.

5.1.- Cross service architecture

User data might be shared to support very different use cases from different clients, and eventually a single dataset may have very different utility and value at the time of supporting different use cases. The PDK must be able to support specific requirements from each vertical but, at the same time, provide the clients with a set of tools and algorithms that help calculate a fair distribution of payoffs first among players in the value chain, and then among subjects providing data that is sold or used by third parties. Moreover, it must be open to the definition of new use cases and verticals that may appear in the future.

5.2.- Flexible pricing

The PDK must support and provide different pricing models in order to match buyers and sellers, based on the ones currently available in data markets. The most widely used pricing models are:

- fixed price set by seller, which provides a description of the dataset and metadata
- price set by a case-by-case bid from the buyer and accepted by the seller(s), sometimes considering a floor price set by the seller

More innovative pricing models shall be studied, which eventually may help bootstrap PIMS and data markets in the economy, by overcoming some of the pitfalls they are suffering from nowadays. For instance, a usual data buyers' concern is that they cannot predict or understand how useful a dataset might be by solely relying on the description provided by data sellers. Multiple flexible pricing models shall be proposed and included in the marketplace design, letting sellers and buyers choose among not only the traditional pricing model, but also among more innovative pricing schemes that reward data and players in the value chain according to the value that a PIMS or data market brings to a specific algorithm or buyer's purpose.

5.3.- Cross-entities revenue split

The PDK must support and provide fair ways to **split the generated revenue between online services (data buyers), human data providers, and any other collaborating entities** between them (e.g. the user PIMS, any data market place, third parties providing data, among others). For example, in the case of a successful sale of a TV set from an electronics' online store, following a recommendation based on purchase data from other users of the store, one would need to decide what part of the price of the TV to return as payback to those users. If the recommendation also used browsing and purchase data of other users without accounts in the e-commerce site (such data can be bought online from so called Data Management Platforms, DMPs), then those users as well should receive part of the payback, as well as the DMP that made their data available to the e-commerce site. Of course, such cross-entity revenue split must consider the contribution of the PIMS to the data value chain.

5.4.- Fair horizontal payoff split

As more than one data subjects will likely contribute to any transacted dataset, it is key that the PDK provides users with fair ways to **compute the relative contribution of different human data providers in creating the outcome of a successful transaction**. After calculating the cross-entity revenue split and determining the payoff to be distributed to data subjects, this functionality is necessary in order to be able to split such payoff in a fair manner among all individuals whose data was involved in the computation.

In some cases, data from different people have more or less equal contribution towards a successful recommendation. That would be the case, for example, of a recommendation for a blockbuster movie, that was produced by counting the number of people that chose that blockbuster movie vs. another one airing at around the same time. In that case, most people that typically watch blockbuster movies would be assigned an almost even share of the amount to be redistributed. On the other hand, if the recommendation was for an experimental Maltese film of the 50's then any users into Maltese cinema should be assigned a much higher proportion than mainstream users watching mostly blockbuster movies or different long-tail context. In a similar way, a driver who reports location data that lead to finding an uncongested route during a peak commute time, or following an accident that blocks a number of lanes, should get a higher return than other users who report back mobility data on already known congested routes.

Enabling compensation schemes that are able to compensate data subjects proportionally to the utility of their contribution is key in order to incentivize the provision of higher quality data. However, the PDK shall provide different compensation schemes so that the PIMS or the data users are able to choose the most suitable one for their business model. For example, one entity might decide to pay for data proportionally to the volume of data used whereas another one might like to provide a variable payment proportional to the accuracy improvement in their AI algorithms yielded by each piece of data. Moreover, it may happen that a third company decides to compensate using a payoff scheme that combines both techniques.

5.5.- Scalability

As a consequence of the previous requirements, the PDK must be built upon a **scalable platform** capable of supporting large numbers of users and very high transaction rates.

In its report "European Data Market study measuring the size and trends of the EU data economy", it is stated that "The European data market has increased by over 50% from 2014, when we [the EC] started tracking its value, to 2018 (from €47B to €72B in the EU28)". The overall value of Data economy, which also includes indirect effects in other players and industries, was estimated "to exceed the threshold of 300 Billion Euro in 2018 for EU28, with a growth of nearly 12% over the previous year", 1,17% of GDP. Baseline forecasts estimate a CAGR of the European data market around 6.5%, which is x4 the GDP growth estimates (i.e. 1,6%) for the same period.

In conclusion, data market is expected to grow in relevance and be one of the key catalyzers of digitalization in the society, and it is consequently a key subject matter within EU policy-making processes within Digital Single Market. Consequently, the design of the PDK must

take into consideration that it will have to provide a near real time service for a huge number of data transactions for the UCDE to be successfully bootstrapped in the economy.

5.6.- Transparency and attestability

For a UCDE to have a realistic path towards adoption, *establishing trust will be key*. Users need to be able to trust that payments in exchange for their data are fair, and in accordance with the payment schemes offered by different services. To achieve this, **payments need to be *transparent* and *attestable***. Transparency and attestability mean that users need to have a way to understand how much and why they were paid for each data use, and do that remotely and in a trustworthy manner that guarantees that the service provider has faithfully followed the advertised compensation plans. Compensation plans can range from very simple ones, like – *pay in proportion to the number of data points provided*, to more complex ones, such as – *pay in accordance to the actual importance of the data provided in the context of a specific computation*, e.g., a recommendation.

Consequently, the PDK must provide tools and functionality in order users are able to understand why they were paid a certain amount of money for their data, and even help users to improve their contributions. Not only must this be done in a user-friendly and easily understandable way so that the community trusts the whole process, but also protecting the privacy of all involved parties, as it is discussed in the next requirement.

5.7.- Privacy

Such **transparency and attestability must be ensured while preserving the privacy of all involved parties**. Particularly, the system must meet privacy preserving requirements regarding:

- The privacy of subjects whose data is managed by the PIMS
- The confidentiality, privacy and intellectual property of data buyers

From a user's perspective, a simple way to achieve transparency and attestability in terms of payoff calculation, is to allow users to see the actual data used in all the transactions carried out by a service, and thus verify for themselves that payments have been fair. Obviously, such an approach is unrealistic. A service has lots of reasons and constraints that make opening up its data impractical, even illegal in some cases (see legal requirements for further information), since beyond jeopardising its intellectual property and competitiveness in its market, it can also harm the privacy of its users (please note here that when we write about opening up its data, we are not referring to allowing a user to see, and even retract his/her data record. We are referring to allowing a user to see the data of other users and thus be able to verify that he got compensated a fair amount. This would break the privacy of users since one would be able to see what websites, or movies, or location everyone else has visited, something that can be intrusive and dangerous to one's privacy).

Similarly, if the system is using data provided by a third party, the PDK must provide tools to preserve the privacy of individual data subjects contributing to that purchased piece of data.

5.8.- Confidentiality and protection of intellectual property

In addition to respecting end users' privacy, **the system must respect the confidentiality, intellectual property, trade secret and privacy of any entity involved in the transaction**. This includes:

- data buyers (e.g. any algorithm or performance metrics provided, or any results obtained by any processing run by the PIMS in order to evaluate the quality of the data provided and that might be sensitive or confidential) and
- any other entity involved (e.g. another PIMS that provides some data or an algorithm to process it) which is providing any piece of information subject to intellectual property rights or trade secret.

This is a key requirement in order to avoid concerns in the industry regarding the adequate and confidential treatment of the knowledge PIMS and data transactions will be built upon. In order to measure the utility of a piece of data for a certain purpose it makes sense to use the same algorithms and performance metrics the data buyer is using in order to forward the most suitable dataset and have a clear view of the degree of improvement provided. For example, in order to optimally calculate what is the utility of data provided to feed a prediction problem, a PIMS will need to use the prediction algorithm developed by the entity that wants to purchase the data. This algorithm might be a core asset of such company and subject to confidentiality clauses. For data buyers to be safe to share such sensitive assets, the PDK must provide a neutral environment in which the PIMS will be able to execute such algorithms while ensuring that confidentiality is not broken and the algorithm is not used in anything else beyond payoff calculations.

In conclusion, the PIMS must preserve the confidentiality of any intellectual property shared by data buyers for this specific purpose, while ensuring that data value is measured using the algorithms, metrics and indications dictated by the data buyer.

Similarly, if the system is using data provided by a third party, the PDK must provide tools to respect any confidentiality and intellectual property agreement signed with such third entities providing data to the PIMS in the wholesale market.

5.9.- Incremental deployability over the existing technologies and services

Last but not least, in order to ease the process of bootstrapping new PIMS using the PDK, it is required that **the solution built is incrementally deployable over the existing technologies and services**. Even though there have been different propositions to change the architecture of the Internet in order to control data provenance and manage content in a better way (see, for instance, the proposition of content-centric networking architectures in the last ten years, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6419703, last accessed January 2020), the adoption of such disruptive solutions would be very difficult since they would require getting rid of years of investment in servers and communication equipment, as well as substantial modifications to the existing applications running on top of the Internet. Consequently, the PDK must assume the current state-of-the-art and Internet paradigm in order to be successfully and easily adopted by the industry.

6.- Technical requirements for the PDK

6.1.- Introduction

The PDK will be designed as a modular system in which different packages, based on the recent advances in technology, will perform specific tasks. The first module, ***Elements to improve data subject privacy (WP2)***, includes functionalities that allow the users to take informed decisions about which information to share, with whom, and where to store it. The second module, ***Mechanisms for the new data economy (WP3),*** focuses on the creation of a transparent, open and easily accessible data market, including tools to valuate users' data and a trading engine. The third module, ***Novel Data Management Tools (WP4),*** includes tools that allow to export/import and exchange data using standard mechanisms and formats, with proper metadata that let the system know data source, data value, and facilitate the data aggregation from heterogeneous sources. The different modules will be made available to PIMS companies in the PDK, so they can test them and try their functionalities.

In this section, we define the technical requirements associated to each of the different modules.

6.1.1.- Methodology for determining technical requirements

The PIMCity consortium integrates partners with substantial background and large expertise in the different technical areas that are required to design and implement the proposed modules, as well as for the testing and assessment. For instance, NEC and POLITO are experts in the development of machine learning algorithms needed to analyse and extract value from the data. Moreover, LSTECH and POLITO will bring the big data experience necessary to ensure the scalability of the proposed solutions. ERMES will bring the cybersecurity experience, while FW the Cloud experience and IMDEA their wide knowledge of data economy. Furthermore, the development of privacy preserving solutions is well covered with UC3M, POLITO and NEC. Therefore, the technical partners of the PIMCity consortium possess all the necessary competences to provide the technical requirements. As such, each of the referred partners is responsible to define the technical requirements for the module(s) they will design and develop during the PIMCity project.

For each one of the aforementioned modules we will proceed with the following methodology:

- A brief introduction of the purpose of the module;
- The description of the system's goals and objectives;
- A high-level overview of the proposed system, which is expected to be completed by month 9;
- The list and explanation of functional and, where relevant, non-functional technical requirements;
- The constrains that technical partners are facing.

6.2.- Technical requirements for the elements to improve data subject privacy (WP2)

6.2.1.- Introduction to WP2

The PIMCity project is composed of many software modules that concur at building the PDK. To allow scalability and flexibility, each module is an independent piece of software that implements a single functionality. All modules are designed and will be implemented to cooperate harmoniously with each other to build the whole EasyPIMS system. The WP2 defines the architecture and design of the modules that are in charge of protecting users' privacy. It defines the tools and the set of techniques to be used to store safely and process users' data. Moreover, it defines the Personal Privacy Metrics that will be used in EasyPIMS' Transparency Tags to communicate to the users which information each service is collecting in a form easy to understand.

To this end, we define four main building blocks that implement the required functionalities. The first two modules, named Personal Data Safe (P-DS) and Personal Privacy-Preserving Analytics (P-PPA), ensure that users' information is stored and processed in a privacy-preserving manner. The Personal Consent Manager (P-CM) ensures that the user has full control over the data fed to the system. Finally, the Personal Privacy Metrics (P-PM) provides easy to understand metrics to inform the users about which information each system collects.

6.2.2.- Goals and objectives

The aim of the module is to investigate tools to improve end-users' privacy. Machine Learning algorithms will be investigated to define, quantify and control data being collected by online systems. Among the objectives on the topic we have:

- Design and develop a system able to empower the users to control their consent settings inside a platform. It should have an easy to use interface and provide auto configuration options to make it easier for the users to configure complex scenarios by using aggregated/crowdsourced data of the different users to build a set of common profiles.
- Design the Privacy Metrics to i) unveil and communicate to end users the data collected by online services, ii) automatically identify and pinpoint possible privacy violations in data collection, iii) deliver these findings to the end users with an easy and intuitive user interface.
- Develop a set of reusable modules for data anonymization which will allow to analyse users' data without affecting their privacy. It will offer some algorithms and methodologies able to provide a certain level of anonymity using concepts as zero-knowledge proof or k-anonymity.
- Design and develop a system that allows to store the data safely and securely, including the primitives needed to ensure a smooth operation in different scenarios and to contemplate the very heterogeneous types of data PIMS may need to store.

The design of this storage system will be based on open-source solutions such as Mongodb[122], Mysql[123] or HDFS[124].

6.2.3.- Proposed tools

The WP2 includes the design and implementation of 4 software modules, described below.

Personal Data Safe (P-DS)

The Personal Data Safe (P-DS) is the means to store personal data in a controlled and reliable form. It implements a secure repository for the user's personal information. It is responsible for storing and aggregating user's information such as navigation history, contacts, preferences, location history, personal information, etc. This can be designed in Push or Pull mode, i.e., the user can actively decide which information to store and retrieve; or the system can do it automatically by importing information as they are collected while the user performs his usual activities like browsing the web or moving in a city. The P-DS can store either the original copy of user data or point to other repositories, e.g., to external services that have already collected the data. For the design of the P-DS we will adopt the state-of-the-art platforms for web service management such as Django, Flask and Bottle. The service will offer web-based APIs for authentication, data insertion, update and deletion. It will also offer a minimal web interface for the management of data directly by users.

Personal Privacy Metrics (P-PM)

Personal Privacy Metrics is the means to increase users' awareness about data exposed to services. For each service, this module collects, computes and shares easy-to-understand novel privacy metrics, based on which information service is collecting, how it stores and manages the data, if it shares it with third parties, etc. All these pieces of information are fundamental for a user to know and to take informed decisions. The P-PM computes and offers this information via a standard interface, offering an open knowledge information system which can be queried using an open and standard platform. We will use state-of-the-art methodologies to automatically compute the privacy metrics, using supervised machine learning and artificial intelligence approaches, where domain experts, volunteers, and contributors collaborate to populate the information. For the design of P-PM we will consider technologies which build on dynamic code analysis (e.g., OpenWPM) for automatically browsing millions of web pages and analyse code in them, as well as static code analysis to identify latent data-collection routines impossible for dynamic analysis to detect. We will then classify pages and web services based on the information they collect and generate scores to summarize privacy risks. Results will be then published in a database made accessible to other elements of the platform, and periodically revisited and refreshed to guarantee maximum accuracy. Services and data buyers participating EasyPIMS will be provided the means to comment and provide feedback on their results using web-accessible forms.

[122] https://www.mongodb.com/
[123] https://www.mysql.com/
[124] https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Personal Consent Manager (P-CM)

Personal Consent Manager is the tool by which users will define the permissions for their data to be available and processed on the platform. Users, the owners of the data, will have total control to specify what data can be traded and under which conditions. The P-CM should be easy to use for non-expert users but at the same time granular enough to be effective. To that end, we will provide a set of templates of common use-cases (plans) so that users can subscribe to. Plans can be shared to facilitate re-use and mitigate the burden of fine-grained configuration to most users, especially the less tech-savvy. A finer control of permissions will also be available, by each data type can be assigned an availability level (non-tradeable to fully-tradeable) with certain conditions on the trades. For instance, the user might define that despite a particular data-element is tradeable, she will not agree on the trade if certain conditions are not met, for instance, price too low, declared purpose of data to be acquired not desirable (purpose), not enough users participating on the trade (lack of quorum) or timespan on which the consent is applicable.

Personal Privacy Preserving Analytics (P-PPA)

These are the means to impose control of personal data. Privacy is a must. When exchanging pieces of information with systems, we need the ability to know and control which data we are exposing. Concepts like Zero Knowledge, Differential Privacy, K-Anonymity are well known in the cryptography and privacy community. PIMS must offer and leverage these capabilities so that data can be exchanged among different systems while preserving the actual information as private. The P-PPA will offer standard and open implementation of these fundamental methodologies. We build on state of art, and commoditizing it so that any PIMS based on the PDK can easily integrate privacy-preserving analytics. We will provide open implementations of the popular algorithms for anonymity such as k-anonymity, l-diversity or t-closeness as well as custom approaches to anonymize personal data collected in a web environment, with particular reference to browsing history of users. All techniques will be implemented to ensure scalability when possible, and using state-of-the-art frameworks for analytics such as Apache Spark and MongoDB.

6.2.4.- Functional requirements

Requirement Name	Mandatory/ Obligation/ Recommended	Description	Elements Involved	Notes
R1: Login	M	Users shall be able to login in the system prior to the self-assessment. Login can be done automatically by other modules (e.g., P-CM).	P-DS	Actor: Users, Data Buyers, others system elements

R2: Insert/Modify data storing preferences	O	Users shall be able to configure which data the P-DS stores, e.g., browsing history or location.	P-DS	Actor: Users, others system elements
R4: Consult the data users provided to the system	O	The users shall be able to consult the data stored in the system in an easy way	P-DS	Actor: Users
R5: Query the user's data in the system	O	The data buyers shall be able to run flexible queries over the data for which the users provided their consent. The P-PPA guarantees that users' privacy is respected.	P-PPA	Actor: Data Buyers, others system elements
R6: Login	M	Routines responsible for generating Privacy Metrics shall be able to login automatically in the system to insert new Privacy Metrics or update existing ones.	P-PM	Actor: Other system elements
R7: Insert/modify per-service privacy metrics	O	Privacy metrics will be generated by routines processing web services data generated by automatic crawlers or traffic logs. Routines will calculate privacy risk/scores based on data collection patterns identified in the code, the nature of collected data and connections to third parties.	P-PM	Actor: Other system elements
R8 Consult per-service privacy metrics	O	Users shall be able to consult data (reputation, privacy scores) about web services	P-PM	Actor: Users

R9: Query per-service data in the system	O	System elements shall be able to query data (reputation, scores) about web services	P-PM	Actor: Other system elements, Users
R10: React when per-service data in the system change	R	System elements shall be able to react to changes in data (reputation. scores) about web services	P-PM	Actor: Other system elements, Users
R11: Login	M	Users shall be able to login to define their consent policy / permissions.	P-CM	Actor: Users
R12: List Active Consents	O	Users shall be able to see all the consents active on their data. By default, unless a consent is created all data is not subjected to trading.	P-CM	Actor: Users, Other systems
R13: List All Available plans	O	Users shall be able to see all the plans (aggregations of consents) available to the public.	P-CM	Actor: Users, Other systems
R14: Create New Consent	O	Users shall be able to create a consent for a data element or type, with different constrains such as: time-to-live, price, purpose, etc.	P-CM	Actor: Users
R15: Remove Consent	O	Users shall be able to remove constrains.	P-CM	Actor: Users
R16: Create New Plan	O	Users shall be able to create a plan (a set of individuals consents) and make it public.	P-CM	Actor: Users

R17: Subscribe to Plan	O	Users shall be able to subscribe to a plan, which will add the individual consents that belong to that plan.	P-CM	Actor: Users
R18: Unsubscribe to Plan	O	Users shall be able to unsubscribe to a plan.	P-CM	Actor: Users
R19: Query the system	O	The data buyers shall be able to make queries to the data, only using those fields for which they have consent, guaranteeing users' privacy.		Actor: Data Buyers

6.2.5.- Constraints

The design of the software modules for WP2 will respect several constraints to ensure a high-quality for the software architecture, code and security operations.

Several WP2 modules include functionalities for user authentication (e.g., login and logout). Moreover, the modules will interact using web-based APIs. As such, it is necessary to ensure that communications are secure and users authentication is robust. To this end we will leverage state-of-the art tools for secure web-based APIs. All communications will be encrypted using HTTPS and we will make use of authentication tokens to enforce control on the use of the APIs.

Three WP2 modules, namely the P-DS, P-PPA and P-PM, need to operate with a large amount of data, potentially from millions of users. As such, these modules need to be scalable and designed to be deployed on a distributed architecture, so that they can handle small to large communities of users with minimal effort. To this end, we will use scalable-by-default tools, such as Django, MongoDB, Spark.

Finally, all the phases of development must follow the best practices for development. We will follow the agile software development approach and the DevOps paradigm for automated build and test, continuous integration, and continuous delivery originated. All the code will be open source, and available on some popular repository that will be chosen later in the project (e.g., GitHub or GitLab).

6.2.6.- High level architecture

We here provide a high-level schema describing the interactions of the WP2 modules each others and with the other elements of the system.

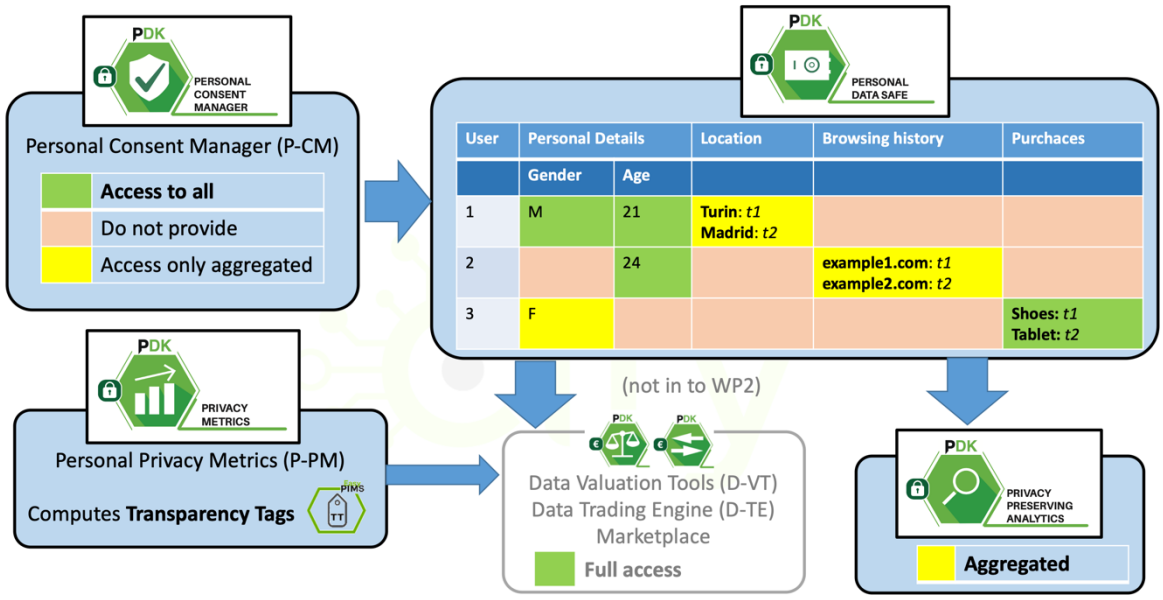

Figure 8: High level architecture of WP2

6.2.7.- Interfaces

In the following schema we illustrate the interfaces of the WP2 modules with the various PIM stakeholders. The modules shall interact with users as well as data buyers and researchers.

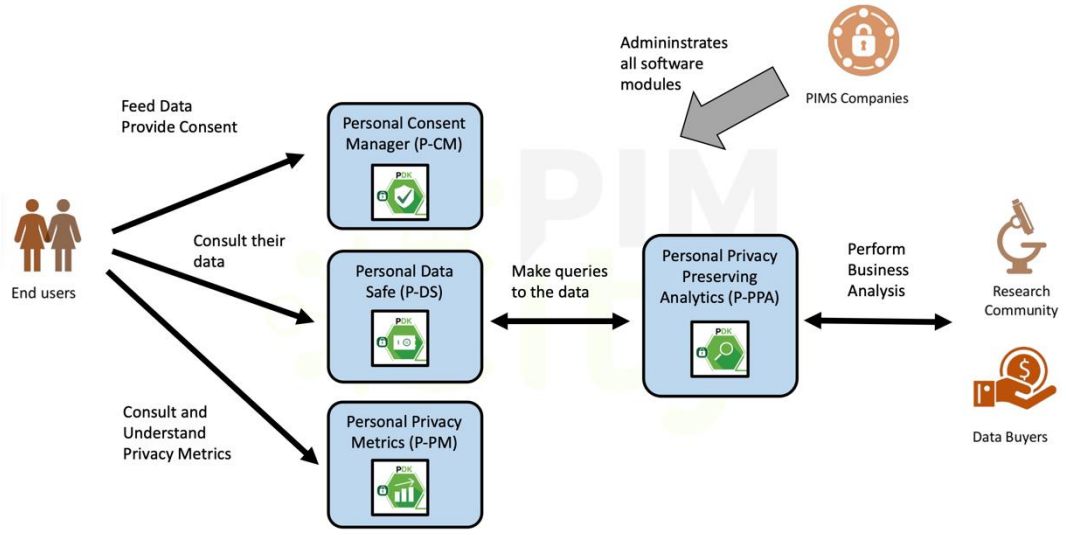

Figure 9: Interfaces of WP2

6.3.- Technical requirements for WP3

6.3.1.- Introduction to WP3

The definition of a new human-centric data economy which provides high quality data for businesses and at the same time respects the privacy of end-users is a must. The first mission of this WP is to provide a detailed model of such new human-centric data economy model. This specific aspect of the WP has been treated independently in this deliverable and all details related to its requirements can be found in Section 5.

As a second mission in this WP we try to address a long-standing question largely unknown in the current data economy ecosystem: *what is the value of end-user's data?* This is a very challenging issue to address, which is key for the development of any successful PIM. End-users' data is the main asset PIMs are going to trade with, thus, knowing its value is of paramount importance to develop an appropriate business plan which assess the profitability of the business run from each specific plan. Currently, users are not part of the data market, they are external agents who just provide the asset but do not have any influence or decision power. In this scenario, the value of end-users is solely defined by the market, i.e., what the data buyers are willing to pay for the data of a specific end-user. However, in the PIMCITY's foreseen human-centric data economy, this one-sided vision is not valid anymore. In the human-centric data economy, end-users will have control over their data and will have the last word on what they are willing to share and with which third parties. Hence, we come into a new scenario where we have two sides: the market and the users. Each one of them will have its own view on the value of end-users' personal data. One the one hand, market will define the value based on the demand vs. offer theory, while end-users will have its own perception which may involve many different aspects (privacy, social status, education, etc.). In this WP we would develop technology able to infer the value of end-users from the two mentioned perspective: market vs. end-user perspectives.

Finally, the third commitment of this WP is related with the development of what we refer to as a *trading engine* technology. This is a piece of software which can be integrated as part of the PIM infrastructure, allowing to trade end-users' data within the data ecosystem. The *trading engine* will be designed to be integrated as part of the PIMCITY's *Data Broker*.

6.3.2.- Goals and objectives

This WP defined a list of tangible objectives that we describe next:

- Define the entities involved in the data market.
- Define the interaction between the different entities participating in the data market considering solutions of different nature ranging from centralized ones to smart contracts based on distributed blockchain.
- Implement tools to provide estimations (from the market and the user perspective) of the value associated to individual users' data.
- Design and implement a system, referred to as the trading engine, that integrates the data broker into the data market ecosystem allowing it to sell end users' data to interested third parties.

6.3.3.- Proposed tools

Data Valuation Tools from market perspective (DVTMP)

In the last decade, complex tracking infrastructure and ecosystem has been developed on top of the Internet and the global telecom networks. This tracking infrastructure leverages both the web and mobile ecosystem, so it cannot only track our online activity (what websites we visit or which products we buy online) but also our offline activity (what places a user goes can be extracted from her mobility data obtained, for instance, through mobile apps).

This tracking infrastructure serves to collect data from users and profile them. Arguably, the most relevant business driving this tracking ecosystem is online advertising. Companies, can know what are the habits and preference of individual users and thus they can deliver personalized advertisements or offers to each user.

Many of the most popular online services follow this business model, i.e., they define different techniques to profile users and provide them with personalized ads, products, offers or services. Among them we find Google, Youtube, Facebook, Instagram, LinkedIn or Twitter. Most of these companies provide web platforms or APIs to allow advertisers defining their advertising campaigns. In many cases, part of the information they offer to advertisers through those interfaces includes an approximation to the end-user value. In particular, they offer advertisers estimations of the Cost Per Mille (CPM), Cost Per Click (CPC), etc. Hence, when an advertiser configure its target end user profile (a.k.a. audience) in the service platform (e.g. Facebook platform) it will get in return the cost of showing 1 thousand ads to users meeting such profile (CPM), the cost of getting a click in the ad bring the user to the advertiser's landing page (CPC), etc.

While the aforementioned services operate as standalone online advertising providers, there is also an open market where ad spaces in webpages and mobile apps are sold based on an auction process. Let us consider the following example: a user U visits a webpage (e.g., acme.com) which includes an adspace. This adspace is offered in an auction to advertisers. This auction provides advertisers different types of information: 1) **ad space info**: the type of ad space (banner, video ad), the location of the ad space (top of the page, bottom of the page); 2) **website/mobile app info**: what are the categories that define the website/mobile app owning the ad space; 3) **audience information**: profile information about the user exposed to the adspace. Based on these data interested advertisers place a bid for the adspace. The intermediary entity running the auction process (referred to as Ad Exchange) choses a winner for the auction (e.g., the highest bidder) and information about the price of the ad is exchanged between the parts. All this occur in real time (few hundreds milliseconds) and it is orchestrated through the OpenRTB protocol. Hence, the price of winning bids would provide a good estimation about the market value of the specific audience (i.e., user's profile) associated to the auction process.

Therefore, the Data Valuation Tools from market perspective (DVTMP) developed in PIMCITY will leverage the existing online advertising platforms and protocols to obtain an estimation of the value of hundreds to thousands end-users' profiles representing an important portion of the population. This will meet the specific objective defined in this WP in this regard.

Data Valuation Tools from end-user perspective (DVTUP)

Once data has been shared and an economic value has been generated by the marketplace, how to split such a value between the entities and parties contributing to such

data transaction becomes a relevant problem. This problem may be further decomposed in two:

1) Splitting such value between the different entities contributing to the transaction, namely the PIMS or data market, the users providing data and / or third parties contributing with their data.

2) Splitting the value among end users that consented to provide data for such transaction.

In general, the solution to the first problem depends on market forces and the **structure of data value chain**. Consequently, it is important to present the results of the benchmark in the context of a commonly shared data market structure, which will also be elaborated as a valuable tool for the whole project. A number of different PIMS and data marketplaces already provide data market services for different purposes, from user profiling for marketing campaigns, to providing data for ML algorithms. A **comprehensive PIMS, data provider and data marketplace data benchmark** will be performed in order to answer questions such as:

- Which kind of data is it intended to manage? Is it a general-purpose marketplace? Whose data? Individuals? Enterprises? Both?
- What does the data buyer get out of each transaction?
- Who determines the structure of data to be stored? Where is data stored?
- Does the PIMS / Data marketplace combine data from their data owners with other data from third parties?
- Who pays whom? What is the business model? What is the percentage of value
- How is data sold in the market? Who establishes the price? In accordance to what?
- How does the marketplace redistribute payments for data?
- How can users know who used their data and for what? How can users know who used their data and how much they earned because of it?

The main objective of such benchmark is to better understand the **business models** and how companies currently operating in the market are approaching those two problems. Notwithstanding the above, it will provide relevant insights into the current data trading markets that will eventually help companies entering the market design novel models.

Nevertheless, additional tools will likely be needed to help PIMS and data marketplaces split the economic value among the different data sources that were part of the transaction. The DVTUP module will implement a series of tools and algorithms that will facilitate this task for PIMS / Data marketplaces using the PDK:

- **Payoff distribution techniques based on contribution metrics** from the different users, which will allow to i) collect and feed some metrics, ii) calculate the value split of a single transaction or group of transaction among the users of a PIMS, iii) record all the required information to provide transparency and accountability of the described process. This may include, for instance, a payoff distribution according to the volume of data provided to the marketplace, proportional to the number of times that a user's personal information is shared with third parties, or both.
- As it has been extensively proved that not always the demand volume is necessarily a good proxy of data value even in the context of a single transaction, we will develop a specific set of tools to compute **payoff redistribution techniques based on the value** that the different information sources bring to an algorithm. We will resort to

Shapley value from collaborative game theory as a baseline metric for establishing the importance of data sources and individuals in the context of a coalition of data providers. The Shapley value has many salient fairness properties and wide market adoption, but at the same time entails serious combinatorial complexity challenges since its direct computation in a coalition of size N requires enumerating and calculating the value of $O(2^N)$ sub-coalitions. We will provide different algorithms and implementations that allow PIMS and data marketplaces to approximate the Shapley value in polynomial time in a general context.

In addition, we will support both the possibility that payoff distribution is calculated and accounted for each transaction, and that such redistribution is computed and accounted for a number of transactions or over a period of time (e.g. monthly redistribution of fees received by the PIMS for the use of personal data).

Trading Engine

The Trading Engine (D-TE) must allow buyer of data to acquire data of sellers (users) with a given set of constrains. The D-TE will be responsible to do the matchmaking between requests (contracts) and the assets available (offered by the users/sellers).

Sellers can offer their data, stored in the Personal Data Store P-DS, according to the consent defined on the Personal Consent Manager (P-CM).

The data offered can come from a wide variety of sources: 1) Explicitly given by the user (via forms), 2) Automatically derived (for instance, categorization of interests based on browsing history), 3) Bulk data (institutions could upload data on their platform that belong to a given user. For instance, a user could ask a telco to periodically upload location data of their users, or could ask Facebook to upload the user's profile it automatically generates.

Regardless of the source of data, all data available will be stored and managed by the P-DS, and the permissions and consent managed by the P-CM. All data-assets must be labelled on a **taxonomy** which should be extensible. For instance, some examples follow about possible tags and the kind of information they should be leading to:

tag:location:city => 'Munich'

tag:location:current_location => GPS coordinates, last timestamp

tag:location:history => Zip file with GPS locations with timestamps

Besides data-assets, the taxonomy should account for attributes that would allow user segmentation, which could be accomplished by using industry standards such as Google Ads and/or IAB taxonomy for user's classification. One of the outputs of the project would be such an extendable taxonomy.

On top of data, which can be used for user segmentation or as the asset to be traded, the system should also account for **attention**, which can be summarized as the commitment of the user to receive a piece of information (most likely an ad) instead of sending data. Basically, users not only can they offer data (in whatever format) but also the willingness of receiving ads. This way, we can use the 'same system' to satisfy the 2 most important use-cases we foresee:

1) Buyers interested in raw data (based on some segmentation),

2) Advertisers interested on accessing users (also based on segmentation).

The TE module should allow sellers to propose contracts (via API or Web Interface) on what data they are interested based on certain constrains. For instance, a contract might contain,

tag:location:city == 'Munich' and tag:attention == true

which should return a set of users who are in Munich and are willing to receive ads. Or,

tag:location:city == 'Munich' and tag:location:history

which should return a zip file combining all location history of people in Munich. Of course, that only affects the users who have given the content for this particular data to be accessible through the P-CM module. For instance, a user, through the content manager might give the consent to access t*ag:location:city* but not *tag:location:history*.

The contract should also define,

- Price scheme, either explicitly by seller a and/or buyer or given by the DVTMP module
- Protocol (data is automatically accessible if conditions are met, or data is held for the explicit consent by the user)
- Institution
- Purpose (also to be defined on a taxonomy). Users should have the ability to opt-out of certain purposes and institutions.
- Other to be defined.

Once the buyer submits a contract (a query), the trading engine will evaluate against all the data according the constant and additional constrains and return the pointers to the actual data or pointer to the user's attention. The trading engine also should account that some pointers might be help by the user's final agreement if specified.

All contracts and subsequent transactions must be logged for auditing purposes. The trading engine (and PIMCity) acts as a trusted party.

6.3.4.- Functional requirements

Requirement Name	Mandatory/ Obligation/ Recommended	Description	Elements Involved	Notes
R1: Crawl data value of audiences from the selected advertising platforms and protocols	O	The DVTMP must provide a function to collect the estimated value of audiences (i.e., users profiles) from	DVTMP	Actors: Online advertising platforms, Crawling submodule
R2: Process, clean and	O	The collected raw data must be processed and	DVTMP	Actors: Data Curation, process

curate the collected data		cleaned appropriately, discarding the incorrect and corrupt collected data and formatting the data for its storage		and storage submodule
R3: Store processed data	O	Once the data has been processed and formatted it should be stored in a systematic manner, e.g., using a database.	DVTMP	Actors: Data Curation, process and storage submodule
R4: Data Value accessibility through an API	O	The access to the formatted data including value of audiences (profiles) available at the DVTMP data repository will be accessible through an API.	DVTMP	Actors: Interfacing submodule
R5: Data Value accessibility through a web interface	R	To grant access to the data in the DVTMP module to not skilled users, it should implement an intuitive web interface for this purpose.	DVTMP	Actors: Users, Interfacing submodule
R6: Automatic error detection	R	It would be desirable that the DVTMP implements an error detection tool for each of its submodules so errors can be promptly identified and appropriate actions for fixing them are taken.	DVTMP	Actors: All submodules
R7: Crawler information up-to-date	R	The value of audiences fluctuates as any other asset in a market. We should then keep the value of audiences updated	DVTMP	Actors: Crawling submodule We will analyze the dynamism in the change of data value in each platform. Scalability,

				R7 and R8 may be subject to a trade off
R8: Crawling speed control	R	Some platforms may impose a limited number of requests. If this is the case we need to adapt the crawling speed to it.	DVTMP	Actors: Crawling submodule We will analyze the dynamism in the change of data value in each platform. Scalability, R7 and R8 may be subject to a trade off
R9: Data buyers to provide valuation functions and/or algorithms	M	Data buyers opting for data pricing schemes based on the value that data brings to their algorithms must provide such algorithms, as well as a valuation function or method to check the accuracy of the results obtained.	DVTUP	Actors: Data Buyers, D-TE, DVTUP
R10: Track data used in transactions	M	DVTUP module shall be able to track the data potentially usable or actually used in data transactions according to consent parameters of the end users, as well as it characteristics (e.g. volume)	DVTUP	Data shall be brought from the P-DS and data taxonomy provides a way to classify and track useful data for a transaction
R11: Calculate distribution of value based on simple drivers	M	DVTUP module shall calculate volume-based distribution (or either using other simple metrics of datasets) of value in a transaction.	DVTUP	Internal functionality
R12: Store the results of transaction value billing	M	DVTUP module shall be able to store the results of transaction value calculations so that the PIMS can	DVTUP	Actors: DVTUP and D-TE, who will provide the value of such transaction

		use them to potentially redistribute payoffs or rewards data sources		
R13: Calculate value-based distribution of the value of data transactions	M	DVTUP module shall calculate the value that each individual data source is able to bring to a transaction, based on the value that each source brings to the traded dataset.	DVTUP	Internal functionality
R14: Performance issues	M	DVTUP module must be able to evaluate the high volume of transactions that are expected to be carried out by the data trading engine.	DVTUP	Internal functionality
R15: Accuracy vs Speed control	M	DVTUP module shall provide the users with algorithms with different accuracy vs. computation time tradeoff, and let D-TEcontrol which of them will be used or automatically control the algorithms applied depending on the system load.	DVTUP	Actors: D-TE and DVTUP
R16: Accessibility through APIs	M	The module shall provide interfaces to the PDS, user dashboard and data trading engine to be integrated in the full data trading process.	DVTUP	PDS used to retrieve users' data. User dashboards to provide information that helps provide transparency to the process D-TE to get information on transaction and model or algorithm used value and provide results
R17: Provide an execution environment for	M	The module and the underlying system must provide an	DVTUP	Internal functionality

the buyers' algorithm		execution environment for the algorithms and valuation functions provided by data buyers		
R18: Store additional information to ensure transparency	R	The module shall store more detailed results of the calculation in order to ensure the transparency of the process and be able to report to each customer about their specific data value.	DVTUP	Internal functionality. E.g. interim results of the calculations to compare with average users that also provided information for such transaction.
R19: Recommended algorithm to use	R	The module should provide a recommended algorithm to use depending on the specific problem, time and scalability constraints.	DVTUP	Internal functionality
R20: Data Buyer Authentication	M	The Data Buyer authenticates on the TE to start operating	TE	Actors: Data Buyer, TE login module
R21: Explore	M	The Data Buyer queries the TE to test if there are sufficient Data Sellers to buy data from for the desired data type and for the given segmentation parameters	TE	Actors: Data Buyer, TE user interface, P-DS
R22: Offer creation	M	The Data Buyer creates a Draft Offer on the TE to buy Sellers' Data specifying institution, purpose, price, the requested data type and segmentation data as well as the budget or desired data amount	TE	Actors: Data Buyer, TE user interface

R23: Offer publish	M	The Data Buyer publishes the Draft Offer so that the TE identifies, using R2: Explore, all matching Data Sellers and defines the strategy to transact with each one according to his D-CM configuration	TE D-CM	Actors: Data Buyer, TE user interface, D-CM
R24: Notify Data Sellers	M	Triggered by the strategy for Data Sellers that require explicit consent the TE notifies all matching Data Sellers about the recent published Offer	TE	Actors: Data Sellers, TE user interface
R25: Receive notification	M	The Data Seller receives the notification regarding the intention of the Data Buyer to buy his data	TE	Actors: Data sellers
R26: Accept offer	M	The Data Seller indicates that he wants to sell the required piece of data to the Data Buyer notifying the TE about it	TE	Actors: Data sellers, TE user interface
R27: Reject offer	M	The Data Seller rejects the Offer notifying the TE about it	TE	Actors: Data sellers, TE user interface
R28: Transact with consent	M	Triggered by the response in R7: Accept offer the TE requests the pieces of information to the Data Seller's DPA and makes it available to the Data Buyer while also ensuring that the Data Seller receives the according value.	TE DPA	Internal functionality. Links to DPA.

R29: Transact	M	Triggered by the strategy for Data Sellers that don't require explicit consent the TE request the pieces of information to the Data Seller's DPA and makes it available to the Data Buyer while also ensuring that the Data Seller receives the according value.	TE DPA	Internal functionality. Links to DPA.
R30: Discard matching Data Seller	M	Triggered by the response in R8: Reject offer, the TE discards the matching Data Seller that rejected the Offer.	TE	Internal functionality
R31: Transaction history	M	The Data Seller accesses the transaction history being able to visualize all Offer parameters and the status of each transaction.	TE	Internal functionality. User dashboards will be accessing to this information to inform data sellers.
R32: Offer progress	M	The Data Buyer accesses to the current state of an Offer being able to visualize how many transactions happened so far.	TE	Actors: Data Buyer, TE user interface
R33: Unpublish offer	M	The Data Broker unpublishes an Offer when he decides to do so, making it unavailable to Data Sellers	TE	Actors: Data Buyer, TE user interface
R34: Unpublish offer by condition	M	The TE decides to unpublish an Offer after a budget goal or amount of data sets is reached, making it unavailable to Data Sellers	TE	Internal functionality. Notified to data buyer.

R35: Segmentation parameters	M	The segmentation parameters used to match Data Sellers should not target individuals. They can't be that specific.	TE	Internal functionality

6.3.5.- Constraints

As described for modules in other WPs, the software modules will be designed and developed respecting several constrains that will provide strong guarantees with respect to the high-quality of the delivered software. In particular:

DVTMP sourcing constraints: The advertisement platform to be used as source of data by the crawling submodule must provide direct or indirect information about audience value

DVTMP scalability Constraints: The DVTMP will be designed to collect and process updated information from hundreds to thousands of audiences (i.e., end users' profiles). This imposes a constrain in the design and implementation of the module since it must guarantee the described scalability criterion. Moreover, the DVTMP API should be designed to handle in the order of thousands to tens of thousands of requests per second. This again imposes a constrain to be considered during the design and implementation of this module.

DVTMP security Constrains: The DVTMP module has to offer state-of-the-art security in both the data storage and the communication between third parties and the module. These high standard security requirements represent a constrain that must be taken into account during the design and implementation phase

DVTMP performance Constrains: The third type of constrain to be considered during the design and implementation of the DVTMP module are performance related. On the one hand, crawling software should be fast enough to collect data at the required speed. On the other hand, the API should provide responses in the range of tens to hundreds of ms. These performance constraints need to be considered in the development of the DVTMP module.

Finally, all the phases of the design and development of the DVTMP software must follow the best practices for software development. We will follow the agile software development approach and the DevOps paradigm for automated build and test, continuous integration, and continuous delivery originated. All the code will be shared and available for the development team on a popular repository (e.g., GitHub and GitLab).

DVTUP functionality constrains: In the case of value-based payoff allocation, we will assume that the buyer will share with the data marketplace the *algorithm* (*A*) which the data is intended to feed, as well as the value metric (*v*) which the buyer wants to optimize. The data marketplace and PIMS will use such metric to measure the value that the different data sellers bring to *A*.

DVTUP scalability constrains: The process of calculating the relative value of a transaction involving the combination of different sources must scale as $O(N^2)$ where N is the number of sources. DVTUP shall offer alternatives with different speed vs accuracy tradeoff to adapt

the accuracy of the calculations to the overall system load. Focusing now on the Trading Engine D-TE, if faces the following constraints:

D-TE matching performance: the D-TE must perform a match between a contract (seller) and a buyer in a matter of milliseconds if the good being requests is attention (which is the typical use-case of online advertisement). An acceptable response time on this feature would be 2 seconds or less to ensure the Explore functionality and Offers take effect immediately. For the system to be credible, it must be up to the speed of real-time bidding systems. In the case that the good subjected to trade is not attention but data, the time-constrains can be relaxed as they are not as time-sensitive. The system, however, must be able to accommodate the worst-case scenario.

D-TE performance: Transactions should take a short amount of time to complete so that the TE is able to perform from ten to one hundred thousand transactions per minute. Moreover, the TE needs to support transactions from many users from different offers so it needs the capacity to handle heavy loaded concurrent scenarios.

D-TE scalability: the D-TE must contain all contracts (which expire on time) but also all offers (which have a much longer expiration, if any). Therefore, special emphasis must be placed on the scalability of the D-TE so that it does not crumble on the scenario of millions of users selling their data.

D-TE security: although the D-TE does not hold any user data on itself, it might leak some information about them unintentionally. Special care must be placed to avoid indirect data-spill by attackers.

D-TE auditing: any agreement reach between buyer and seller must be logged for auditing purposes. It is not the scope of the D-TE to resolve conflicts between both parties, but it must record all data necessary for an out-of-band investigation if required.

6.3.6.- High level architecture

The next figure summarizes the high level architecture of WP3 modules and its interaction with modules from other WPs.

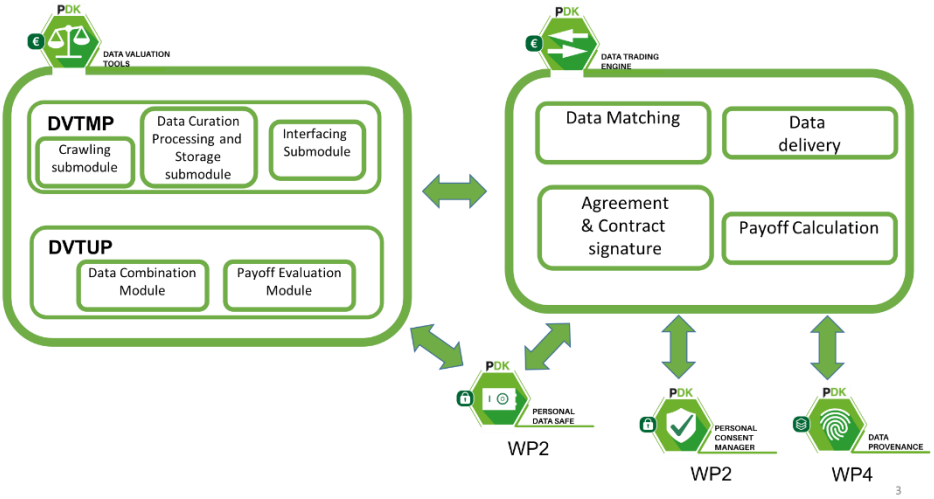

Figure 10: High level architecture of WP3

6.3.7.- Interfaces

The next figure summarizes the interfaces between the different modules developed en WP3 as well as with modules in other WPs.

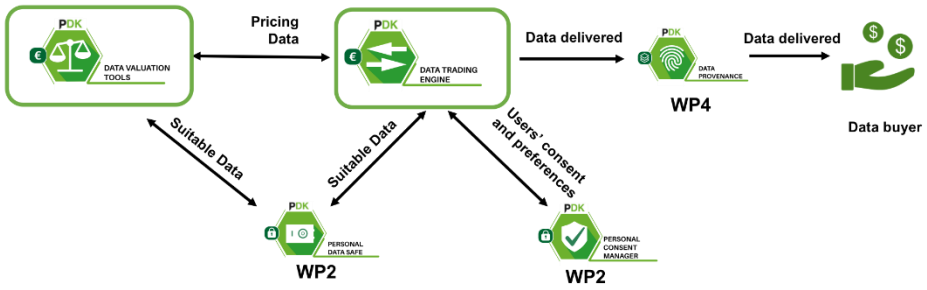

Figure 11: Interfaces of WP3

6.4.- Technical requirements for WP4

6.4.1.- Introduction to WP4

Data management has been considered in different contexts, and especially in the online world, in an effort to empower data producers (in our case, end-users) to be able to manage the data they produce and/or collect. For online end-users, nowadays, data management typically means front-end User Interfaces that are connected to back-end systems and allow them to query said systems for some visuals of the users' data, to gain insights on the data collected. However, such systems rarely extend basic functionalities to the end-user, including modifying or deleting the data, or allowing them to port their data to another system. Furthermore, the end-users typically do not have any knowledge extraction tools in their disposal, for mining their own data, aggregating them in different dimensions, and thus understanding more about themselves. Additionally, if the system collected some data of the user, they can probably share that data with multiple collaborating 3rd parties, and the end-user has no way of knowing who accessed their data, how many times, etc.

Data portability has been proposed in the past few years as a means of interoperability and web openness. In the past, Semantic Web technologies like Web Access Control[125] vocabularies have allowed researchers to model user information and user-generated content, and make it easier to share or port data across platforms. However, these proposals have typically been considered as academic exercises and mostly failed to take off, since there have been no incentives embedded for data processors that operate under information lock-in by building data silos and hoping to maintain their competitive edge against others, to enable data portability.

Interestingly, in recent years, the need for end-user tools and functionalities for better data management, portability, control, storage and sharing, has received new attention, and all such functionalities have been included as requirements for data storing, sharing and processing platforms, as mandated by new EU regulations such as the GDPR, and other recently proposed Internet Policies[126],[127]. Such legislations have practically forced social

[125] Heitmann, B., et al. "An architecture for privacy-enabled user profile portability on the web of data". IHFRS 2010.
[126] B. Engels. "Data portability among online platforms". Internet Policy Review, 2016.
[127] I. Graef. "Mandating portability and interoperability in OSNs: Regulatory and competition law issues", ECTP 2015.

networks (e.g., Facebook) and other online operators to rethink their stance, or face severe monetary penalties regarding data management, control and portability. Thus, now Facebook and Google let the user review/delete their data, and even export their data, but only to basic formats (e.g., JSON or CSV).

In the PIMCITY project, we first take a step back to study past work on data management, portability, control and mining, and understand what privacy threats exist, and can be mounted on users' data during the porting and storing process. The portability process, when successful, effectively allows a user and data owner, to retrieve (all) their data from platform ***A***, and import them, if they desire, into another platform ***B*** that offers similar or better functionalities to the user. During this process, the user's data are vulnerable to various attacks that can expose said data. Therefore, we need to study the range of such attacks that are possible to be mounted, and propose ways to protect the users and their data from them.

Furthermore, and assuming the data are protected from attacks, there is an unmistakable opportunity for the user to review their data when retrieved from platform A, modify them if they want (by correcting, deleting, adding data), and then importing them to platform B. For PIMCity to offer such functionalities to the end-user, it means 1) understanding the type of data at hand that the user wants to input, 2) offering computationally efficient and effective functionalities and tools to process said data (e.g., summarize, select, filter, amend the data, etc.), 3) proposing proper interfaces to allow importing and exporting of data from one platform to another, 4) proposing knowledge extraction tools to better understand said data, including quantified-self tools. Therefore, such data processing functionalities, including functionalities on data porting, aggregation, governance, and knowledge extraction, will enable users to understand better their data at hand, as well as constrain data leakage and control data propagation.

This Work Package (WP4) is entitled "Tools for improved data management", and aims to provide both Personal Information Management Systems (PIMS) companies and non-PIMS companies the necessary tools to ingest and aggregate data in a secure and privacy-preserving way, while offering users data portability and data verification features. Therefore, the target audience of these tools is 1) existing PIMS willing to improve their products and achieve compliance with the GDPR and other legislative frameworks, 2) generic non-PIMS companies interested in integrating specific data management components to improve their products, 3) direct end users who will be provided with a comprehensive set of tools to visualize, manage, validate, track and export their data, and finally, 4) advertising companies which will be provided with more accurate and updated datasets for better audience targeting.

In the next subsections, we explain the objectives of this work package, and outline all the relevant tools to be contributed within this project. Furthermore, we provide detailed explanations of the functional or not requirements, constraints, and basic architectural diagrams and interfaces for each.

6.4.2.- Goals and objectives (TID)

This work package aims to design and develop methods that enable users to perform the following actions with their data:

- Import & integrate data coming from different sources & platforms into PIMCITY

- Aggregate data at different levels of abstraction, to protect user anonymity when desired
- Port / share data to other platforms using interoperable methods
- Maintain control of data usage & access
- Reduce risk of privacy violations from platforms that data are being imported
- Investigate various machine learning methods that will enable PIMCITY users to:
  - Summarize and represent their data into different groups of profiles
  - Automatically extract data value 1) without user intervention, and 2) while being generalizable to different types of user data.

In order to achieve these ambitious objectives, the work package is broken down into four main tasks, as also outlined in the GA:

- T4.1: Design and development of data aggregation pipelines
- T4.2: Generation of tools for easy data portability and control
- T4.3: Design of methods to ensure the data provenance
- T4.4: Design of knowledge extraction methods for user data

Each one of these tasks will produce at least one tool or result that will be integrated with the rest of the PIMCity platform. Some of these tools will also be provided (through APIs and the PDK) to the EasyPIMs platform realization.

6.4.3.- Proposed tools

Data aggregation pipelines

Due to the variety of devices and data sources available, it is a challenging task to import and aggregate data especially in such a way that protects the privacy of the data subjects. At the same time, and to provide valuable information and insights from these data, special aggregation and sharing mechanisms have to be adopted.

The purpose of this task is to define the general mechanisms that will allow the system to import personal data to the PIMCITY platform. In order to achieve that, metadata/ data models should be defined that will allow user data to be stored under a common schema in the platform. On top of these data anonymization methods will be used to preserve users' privacy and security in the data aggregation and sharing processes.

This task will set the basis for the secure data processing activities in the rest of WP4 tasks and the activities are summarized below.

- Define metadata and data models to allow the data import and storage in the platform.
- Define mechanisms for data aggregation and sharing across user devices.
- Define anonymization methods to preserve user privacy and security.
- Compute aggregated user audiences for data trading without exposing users.

The outcome of this task will be the metamodels, the flow and the set of properties of the data to be imported, as well as the functions that can be available and applied to the data in order to allow aggregation in a privacy-preserving way. All the above will be used by the tools that will be developed in the rest of the WP4 tasks.

Tools for easy data portability and control

The purpose of this task is to investigate and propose methods that will allow users to migrate their data to new platforms, in a privacy-preserving fashion. More specifically, it will investigate methods for extracting data from one PIMS (e.g. Facebook, Bank, Mobile Phone), process it by filtering out (e.g., by applying differential privacy) sensitive information such as platform-inferred data (e.g., social interactions between users, or user unavailability due to event attendance) or user-inputted data (e.g., remove login credentials or debit card numbers), and outport it into a new PIMS (e.g., EasyPIMS).

All aforementioned methods for data importing, processing and outporting will be implemented under a software titled Data Portability Control (DPC) tool.

Methods to ensure the data provenance

The data provenance problem is the very general problem of being able to know where a piece of data has come from and who has accessed it in a complex IT system[128]. The data provenance has been discussed since the early days of the web and even before. The world wide web, as we know it, is a system in which solving the data provenance is very difficult, if not impossible. To a large extent this is because the web is uni-, instead of, bi-directional (pages point to other pages but cannot know which other pages point to them) and lacks any trust model (there is no way to trust that any component of the web faithfully executes protocols and mechanisms).

Data provenance is very important for data marketplaces and PIMS for two reasons:

1. privacy, to be able to know exactly who has seen a piece of information, and thus guarantee that the consent rules set by a seller of information have been respected, and
2. economics, knowing who has seen a piece of information, and how many times, is an important prerequisite for deriving fair pricing schemes for data. Next, we will briefly discuss requirements and existing solutions for guaranteeing provenance in different settings.

The methods to be applied depend on the design of the system. In the case of centralized trusted systems, as is the case of PIMS and the model proposed by the PDK, the problem is simple: all it takes is to establish a secure access control either based on Roles or Attributes[129].

Beyond protocols and trusted execution mechanisms, an alternative way to track provenance and have some control over it, is via watermarking. A watermark is in essence a signature embedded inside a piece of information that indicates, without interfering with what the information is to be used for, where the piece of information originated from. Watermarking was initially introduced as a means of protecting copyrighted content such

[128] A systematic review of provenance systems. Beatriz Pérez, Julio Rubio, Carlos Sáenz-Adán. Computer Science Knowledge and Information Systems 2018

[129] RBAC vs. ABAC: What's the Difference? Url: https://www.dnsstuff.com/rbac-vs-abac-access-control

as images and video [130]. Beyond multimedia data, watermarking has been used for protecting other types of data such as Electronic Health Records[131], software packages [132], and databases[133].

Consequently, the data provenance module will be in charge to include a watermark in any piece of data traded by the PIMS or a marketplace using the PDK, once the transaction is finished. Such watermarks must be personalized by buyer in order to track the provenance of any data leakage or unauthorized use of data. Therefore, watermark insertion is part of the dataset delivery process and it must be invoked before the buyer downloads any piece of data from the system in order to make sure it includes the corresponding watermark.

Data Knowledge extraction (DKE)

It is the means to extract knowledge from the raw data. One of the biggest challenges in big data and machine learning is the creation of value out of the raw data. When dealing with personal data, this must be coupled with privacy preserving approaches, so that only the necessary data is disclosed, and the data owner keeps the control on it. The DKE consists of machine learning approaches to aggregate data, abstract models to predict future data (e.g., predict user's interest in recommendation systems), fuse data coming from different sources to derive generic suggestions (e.g., to support decision by users, providing suggestions based on decisions taken by users with similar interest).

---

[130] J. Zhang, A. T. Ho, G. Qiu, and P. Marziliano, "Robust video watermarking of h. 264/avc," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 54, no. 2, pp. 205–209, 2007.

[Watermarking for images and video A. Tiwari and M. Sharma, "Semifragile watermarking schemes for image authentication-a survey," International Journal of Computer Network and Information Security, vol. 4, no. 2, pp. 43–49, 2012.

[11] Watermarking and block-chain for penalizing multimedia breaches. Url: https://eprint.iacr.org/2018/1050.pdf

[131] Privacy by Data Provenance with Digital Watermarking - A Proof-of-Concept Implementation for Medical Services with Electronic Health Records https://www.researchgate.net/publication/221566602_Privacy_by_Data_Provenance_with_Digital_Watermarking_-_A_Proof-of-Concept_Implementation_for_Medical_Services_with_Electronic_Health_Records/link/0fcfd50cbab762276b000000/download

[132] Ensuring Data Provenance with Package Watermarking. Url: https://pdfs.semanticscholar.org/48e4/7f8d1dd3fb6b4b2bae85992dc1ccf20a9be0.pdf

[133] A Comprehensive Survey of Watermarking Relational Databases Research, 2018. Url: https://arxiv.org/pdf/1801.08271.pdf

Watermarking Techniques for Relational Databases: Survey, Classification and Comparison, 2010. Url: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.368.1075

In particular, the DKE will include at least 3 different sets of algorithms that will work in a completely independent way:

- A user data summarization method able to generate a profile of the users interests using raw network data.
- A Quantified self-dashboard and analytics to help users understanding their data (movements, interactions, etc.) and their value.
- A Data taxonomy extraction tool that will provide the user an impression of how the market sees him while also letting third parties to target him under different audiences. As an example, this tool could provide an output similar to what Google shows to its users on the Ad Settings site[134]. This tool will use industry standards and will be based on Google Ads and IAB taxonomy for user classification. The taxonomies list is generated based on the user's activity on the PIMS marketplace, the different data sources that populates the P-DS, rules from the P-CM and also custom user defined tags. The user is going to be able to access this list and also decide whether a system-generated taxonomy represents his interests or not.

6.4.4.- Functional requirements

Requirement Name	Mandatory/ Obligation/ Recommended	Description	Elements Involved	Notes
R1: Include general data attributes	M	The metamodel shall include all the basic attributes of personal data that are available in most common applications.	DA	Personal data of different categories
R2: Support basic and generic data types	M	The metamodel shall be able to support all the major data types that might be used in the various sources.	DA	The metamodel shall accept numbers (various types), strings (in general), strings of specific formats (such as urls, emails, etc), json, arrays etc.
R3: Log the data acquiring process	M	The Metamodel shall store information/metadata about the source of the data, the import process, date/ time,	DA	
R4: Support standard data interchange formats	R	The platform should support the standard data interchange file formats for contact information, like VCF	DA	Standards will be considered and we will support the most common ones. A

[134] https://adssettings.google.com/authenticated

				mechanism that will allow the incorporation of new standards will also be considered.
R5: Data transformation and mapping between the sources and the platform	M	A metamodel for the possible transformation of the source data to the allowed/ supported by the PIMCITY platform shall be defined. Specific mappings from the source data and the PIMCITY data model shall be defined.	DA	Also, the definition of the output types that will be needed when
R6: Support Data characterizatio n	R	The metamodel should allow the characterization of the data attributes in order to support the data aggregation and the privacy functions	DA	For example, we should be able to characterize data with various categories (to be defined) like “sensitive” or “aggregated” or “non-analyzed” etc. This will allow for better filtering, selection and control of the processes/ functions that can be applied to the data, the sharing processes etc.
R7: Data import flow	M	A mechanism that will define the flow of the data import process	DA	All the steps that need to be followed in order for the data to be stored in the platform, including transformations, mappings, aggregations (If needed)
R8: Data privacy levels	R	The import mechanism should support a number of privacy levels that will define the detail of the data to be transferred or the aggregation level etc.	DA	This will be used to define the available actions per dataset and the aggregation/ sharing functions

R9: Data aggregation functions	M	A set of data aggregation functions shall be defined for data types that allow aggregations	DA	Functions like sum, avg, count, categorization etc
R10: Data anonymization functions	M	A set of data anonymization functions shall be defined for specific data types	DA	
R11: Data processing logging	M	The mechanisms defined should keep a record of the data processes applied to the data	DA	The flow and the processes, functions etc. applied to the data imported should be logged
R12: Data import/ export APIs	M	Provide APIs to allow easy and secure data import and export/ sharing	DA	
R13: Authenticate with the Data Portability Control	M	Users shall be considered as authenticated when an authentication already exists in the PIMS platform.	DPC	The Data Portability Control tool will not have a separate authentication layer, instead it shall use the existing PIMS for authentication.
R14: Authenticate with data sources	M	Users shall be able to authenticate with 3rd-party data-sources	DPC	3rd-party data sources are other PIMS (e.g. Facebook, Banks etc)
R15: Customize data import	O	Users should be able to customize the data import process per data source, by choosing the data fields that will be imported.	DPC	
R16: Store data	O	Imported data should be stored locally in an encrypted format.	DPC	
R17: Transform data	O	Users should be able to transform data for removing sensitive personal information.	DPC	
R18: Export data	O	Users should be able to export saved data into one	DPC	

		of more PIMS-based platforms.		
R19: Delete data	O	Users should be able to securely delete all previously saved data.	DPC	
R20: Include watermarks in exported data	M	The system shall include a watermark in all the exported data in order to track information leakages	DP	
R21: Personalize watermarks by buyer	M	The watermarks shall be personalized by buyer and included in real-time once the transaction is closed and before getting the data. This shall allow the sellers to track who leaked the data in case an unauthorized copy is found.	DP	
R22: Support different watermark types	M	The system shall support watermarks for different types of tradable data including databases, multimedia, real-time records, etc.	DP	
R23: Allow for flexible watermarks	M	The system shall support different types of watermarks depending on the size of the transacted data	DP	
R24: Univocally match watermarks to clients	M	The system must be able to securely and univocally match watermarks to buyers in case it is needed to identify who bought a piece of data.	DP	
R25: User profiling	O	The algorithm will be able to generate a profile of the user	DKE Profile	The main function of the algorithm is to generate the profile
R26: Configurable time interval	O	The algorithm should provide easy to configure change on the interval among profile generations	DKE Profile	To adapt to different scenarios, some applications may require a profile for the last month of data while others may need

				information of the near past.
R27: Near Real time	R	The algorithm will be able to analyse the data in near real time	DKE Profile	The algorithm should be able to produce results every 10 seconds (or less).
R28: Profile with the IAB categories	O	The profile should indicate the interest of the users in each of the categories defined by IAB	DKE Profile	It will help with the integration of the system in a real ecosystem.
R29: Profile categories adaptability	R	The algorithm will be configurable to accept different categorizations for the profile.	DKE Profile	It may allow different usages of the technology, for example to detect dangerous usages of the internet, or to enforce parental protection policies.
R30: Data format flexibility	R	The algorithm will accept network data traces in different formats. (i.e., raw data, proxy logs, csv files, ect)	DKE Profile	To allow the analysis in different scenarios.
R31: Login	M	The dashboard shall provide a login function aligned with the global login architecture of the system.	DKE Dashboards & UA	
R32: Compare data shared	M	The dashboard shall easily show to the customers which data is shared with which enterprises, allowing to compare with what other peers are sharing and providing information about average potential income by each data type.	DKE Dashboards & UA	Preliminarily this will focus on personal information for ad targetting and geo-located information
R33: Link to consent management	M	The dashboard shall provide a quick link to consent management console for the users to change their permissions.	DKE Dashboards & UA	
R34: Data usage &	M	The dashboard shall provide a report of transactions over time	DKE	

reward tracking		where data was used and the rewards / payoffs obtained.	Dashboards & UA	
R35: Data reward transparency	R	The dashboard shall provide an explanation about why the user obtained such rewards and value by tapping into the information provided for transparency by the DVTUP module. Such transparency must be provided by safeguarding other users' privacy (e.g. comparison with what kind of info the average user is sharing)	DKE Dashboards & UA	Preliminarily this will focus on personal information for ad targeting and geo-located information
R36: Recommendations to improve monetization	R	The dashboards shall provide the users with recommendations so that they can improve the monetization of their data assets (e.g. companies seeking for data that they are not allowing to share data with)	DKE Dashboards & UA	
R37: Dashboard functions	R	The dashboard should support functions like data export, define custom layout, save layout, filters, link to the source data and their management	DKE Dashboards & UA	
R38: Data analytics console	M	Provide data analytics for user habits based on location and comparison to those of average users, for instance, average distance travelled, areas visited, etc.	DKE Dashboards & UA	This must also improve the transparency of data valuation.
R39: Permission grant	M	The user grants permission for the tool to generate the taxonomies list and share it with the D-TE	DKE Taxonomies	
R40: Data ingestion	M	The TXS needs to be capable of ingesting events coming from changes in D-CM configuration and activity from the Marketplace, and also raw data from the D-PS.	DKE Taxonomies	

R41: Taxonomy generation	M	Triggered by users' new pieces of information, the TXS generates tags that identifies the new data-assets available	DKE Taxonomies
R42: List user taxonomies	M	Allow users to see all the taxonomies the tool has identified	DKE Taxonomies
R43: Remove taxonomy	M	Allow a user to remove a taxonomy he is not willing to share in the market	DKE Taxonomies
R44: Add custom taxonomy	M	Allow users to add a custom tag with a specific value	DKE Taxonomies
R45: Marketplace visibility	M	Triggered by an update on the taxonomies list the TXS shares the information with the marketplace so that the D-TE knows how to segment the user	DKE Taxonomies

6.4.5.- Constraints

The different tools developed within WP4 presents very specific constraints:

The design of the metamodels, flows and pipelines will take into account the limitations of the data sources and their implementations. The communication with the data sources will be based on their exposed APIs or the available data to be exported from them. Security and privacy constraints may be also imposed by the data sources. Moreover, the data types to be imported will be limited to the needs of the use cases and the project and it is not possible to cover all the possible data sources and data types. We will focus on the most popular and well-defined ones. Finally, it is not possible to provide a wide range of anonymization tools/ mechanisms. We will adopt some of the most common and easy to integrate ones.

In the case of the Data Portability Control tool, the connection between the tool and the data sources will be on-demand only and not continuous. As such, the updates from the data-sources will not be synchronous but will be saved to the tool's internal data stores when the user schedules a new data import. Data will only be stored temporarily and in an encrypted format, until the user manually erases the data stores. Moreover, the supported data import methods will be (i) OAuth 2 authorization standard for sources that support such type of authenticated connection (e.g. OpenBanking API), and (ii) manual data upload, required by sources that do not allow automated data extraction (e.g. Facebook). The data exporting process will be focused on supporting other PIMS-based systems (e.g. EasyPIMS). An authentication based on the state-of-the-art authentication standards (e.g. OAuth2) will be

required. The data format will be based on a well-defined data structure such as the JavaScript Object Notation (JSON).

The Watermarking tool will not be able to handle all types of data, so initially the module will concentrate on data that will be used in demonstrators, namely:

- Personal information used for user targeting (e.g. in advertising), as per the taxonomy defined in WP2
- Geo-located information data streams (e.g. position in time of a given user)

In the case of the user profiling algorithm, it will work by analyzing in near real time the browsing patterns of the users, but will not add other 3rd-party sources to the profiling. Moreover, the profile generated will be performed with the online advertising ecosystem in mind, and as such, it will provide a profile based on the categories defined by the IAB.

Finally, the qualified dashboard will support basic visualizations to demonstrate the functionality. Limitations by the available data will be considered. Visualizations and data tables will reflect the granularity of the data. Custom visualizations and layouts will be available only in the extend of what is required to demonstrate the functionality of the use cases.

The design of the dashboard is dependent on the types of data to handle. Initially the module will concentrate on data that will be used in demonstrators, namely:

- Personal information used for user targeting (e.g. in advertising), as per the taxonomy defined in WP2
- Geo-located information data streams (e.g. position in time of a given user)

6.4.6.- High level architecture

We here provide a high-level schema of the architecture describing the interactions of the WP4 modules with each other and with the other elements of the system.

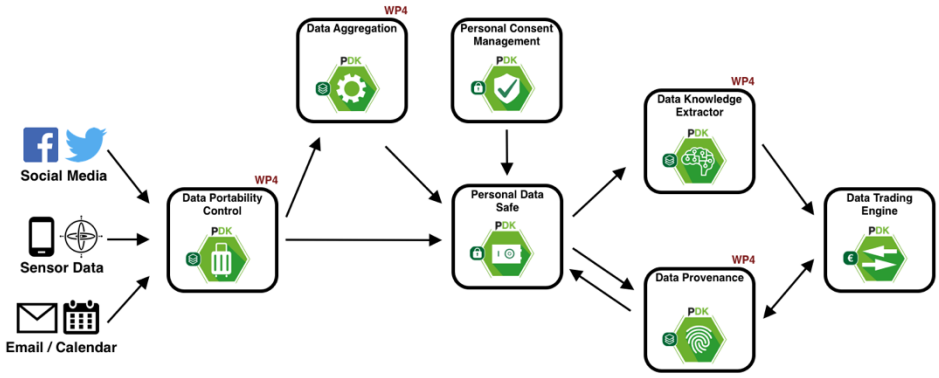

Figure 12: High level architecture of WP4

6.4.7.- Interfaces

In the following schema, we illustrate the interfaces of the WP4 modules. The modules shall interact with end-users, as well as with the available data sources, other PIMS companies and Data Buyers.

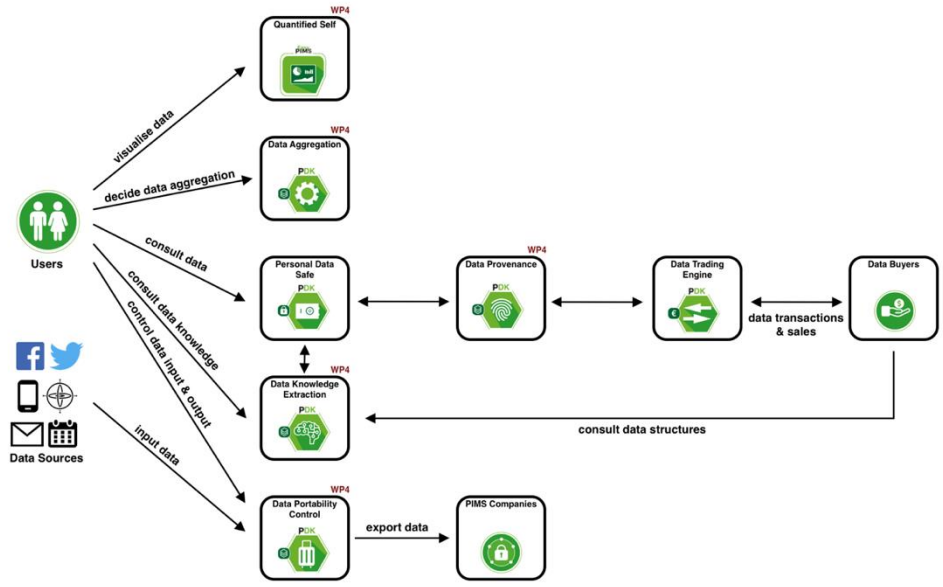

Figure 13: Interfaces of WP4

6.5.- Non-functional requirements

While every component of the PDK presents its own peculiarities, the non-functional requirements can be considered common to most of them. Following, we present a list of basic non-functional requirements that should be met in order to ensure the usability of the solutions.

NFR1: Easy deployment. The different components of the PDK should include documentation and/or tutorials to help non-savvy (i.e., technical people without the highly specialized knowledge to design the different tools.) users deploy the services in different scenarios.

NFR2: Easy to use and understand. The user-facing modules such as the data valuation tools or the transparency metrics, should provide easy to use and easy to understand interfaces.

NFR3: Scalability. The different modules, algorithms and provided APIs should be able to handle at least 100.000 non-simultaneous users. This requirement affects both the sub-systems in charge of the data storage and user data analysis, and those allowing the smooth data trading. Also, the provided APIs may receive thousands of queries per second. Thus, they should be designed and implemented to meet such scalability requirements.

NFR4: Open-Source software. An open-source version of the software will be provided. However, this version will provide only basic features and will not be the full-fledge version.

NFR5: Interoperability and modular design. The different modules should provide interfaces in the form of APIs that will enable module communication in an integrated manner.

NFR6: Accuracy/Performance. The solutions should provide meaningful results. The algorithms will use state of the art technology to achieve the best accuracy possible. The methods that extract knowledge and validate data should not report misleading results.

NFR7: Configuration files. The different modules should provide customization through configuration files in dictionary format that allow users to configure the system conveniently.

NFR8: Logging. The modules should provide logging, monitoring and/or alerting mechanisms to allow the safe execution in production.

NFR9: User Friendly. The different parts that interact with the final users should present user friendly interfaces (UIs), with clean and simple to understand layouts.

NFR10: Time constraints. The different modules should provide results in a reasonable time that allows the execution of different applications presenting time constraints.

NFR11: Legal compliance. All the modules should help the PIMs using them to be compliant with the GDPR and the other applicable legal frameworks.

NFR12: Storage System Security: The system shall provide state-of-the-art security protection to the stored data.

NFR13: Storage System Redundancy and Reliability: Due to the potentially high value of the collected data, including historical data, the system should implement a redundant backup storage system to avoid or at least minimize potential loss of data.

7.- Preliminary requirements for EasyPIMS

7.1.- Introduction

To demonstrate the flexibility of the PDK, we will implement EasyPIMS, a scalable, novel, holistic approach to personal data sharing on the Internet. Key to this is empowering users to concretely understand the value and the nature of data they share. Simplicity is the key to achieve this. For this, we introduce four fundamentals blocks: The Personal Data Avatar (PDA), the Transparency Tags (TT), the Data Marketplace, and the User Dashboard (UD). Each one of these parts will use one or more of components offered by the PDK. Speak about the PDA, TT and Marketplace.

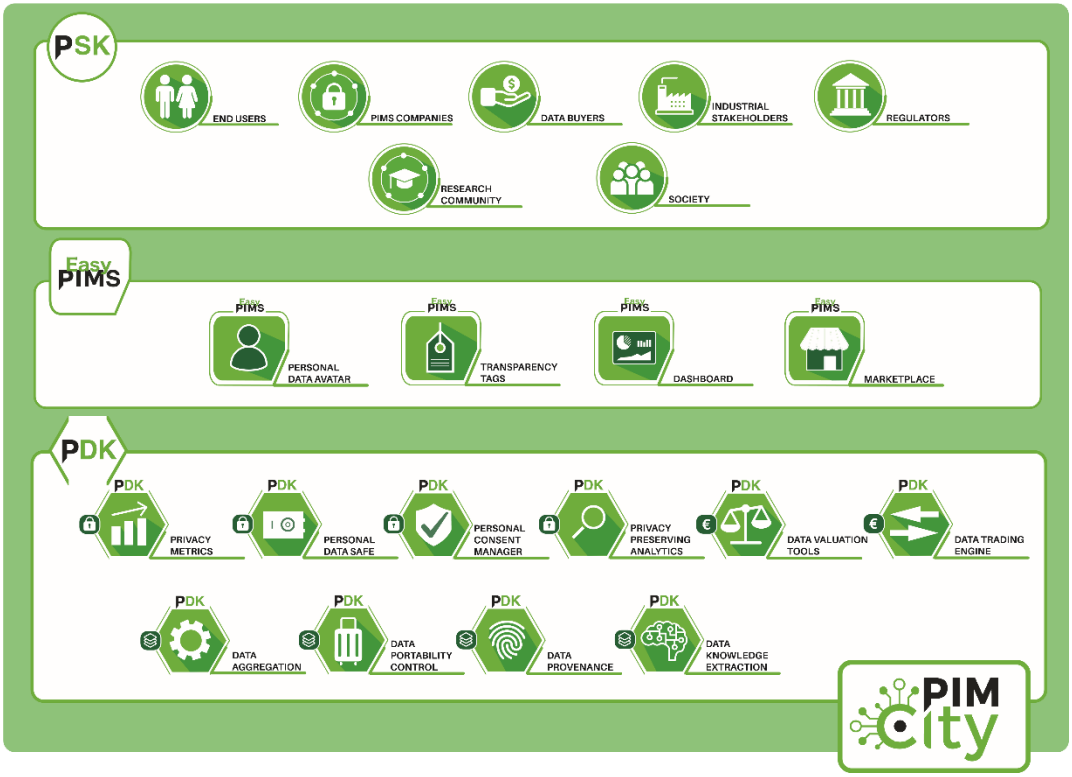

Figure 14: PIMCity components

Figure 14 presents a representation of the different components of PIMCity, including the PDK and the EasyPIMS platform, together with stakeholders which will benefit from them.

We will bootstrap the EasyPIMS platform, feeding it with the data already available at the operators and involving several thousands of real users. PIMCity will leverage the presence in the consortium of two big telco operators with millions of users around the world to start its operation. To this end, EasyPIMS will empower the telco users to share and trade their data in the digital advertising ecosystem.

EasyPIMS will include several elements to allow the trading of data by the users with third parties. Including a ***Personal Data Avatar*** as a representation of the user data; the ***Transparency Tags***, an indication of the privacy lost in the trading of certain information; the ***Marketplace***, where third parties can by the data; and a set of easy to use ***Dashboards*** to allow the managing of the whole system.

The whole EasyPIMS will be developed in the cloud, and a Cloud Controller will be designed and developed to ensure a smooth operation. Moreover, an OpenAPI will be designed and

As in the case of the PDK, we do not only focus on the technical specification, but this document includes the needs of all the parties involved in the data trading, from the users, owners of the data and finally responsible of its use to the telco industry that already stores the data of million of users.

This document presents a preliminary version of the requirements that will be updated in D1.2, after the design of the PDK defines the limitations of the solution offered to the final users.

7.2.- Final users requirements

After the general insights about the privacy concerns of users that was explored in the *Section This* document updates the Version 1 and modify the *Section 2. Final Users requirements for the PDK* and adds a new *ANEX*. In the Section 2 it includes a review of the state of the art about user preferences on privacy, the annex reflects the results of the questionnaire done during the PIMCity project.

Final Users requirements for the PDK. Using the same focus groups, we asked the final users for their idea of a PIMS. To drive the discussion, we asked the users how they would like EasyPIMS to look like.

The menus, usability and design are indicative and are the result of the users' own evaluation of the different options analysed. Thus, they represent the opinion of people that is not professional in the design. Anyway, they will serve as a starting point for the design of the EasyPIMS platform, that will be improved by the technical partners on the consortium.

7.2.1.- Introduction to the UD

The User Dashboard (UD) is the module that connects the end user with the services and applications that, like EasyPIMS, have been designed using the different modules of the PIMCity Development Kit (PDK).

In relation to the UD, in the focus groups, we have worked initially on aspects of usability and structure of navigation.

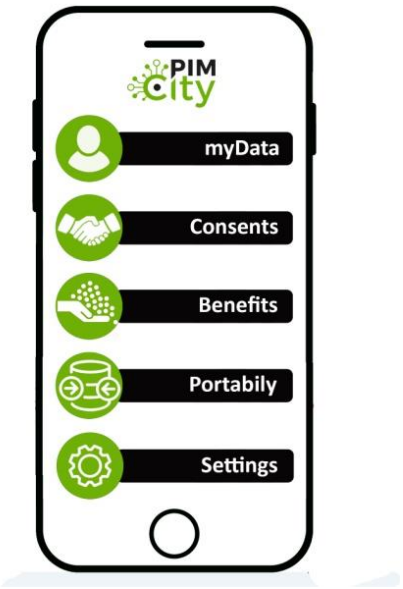

Figure 15: EasyPIMS mobile App mock-up.

The access device most used by users is their mobile phone and therefore the UD must be implemented to make it available for both Android and IoS devices.

The use of this application will incorporate security controls that protect access to it for those who do not have authorization. We describe below the main functions to which the user has access and we have grouped them into five blocks: My data, Consents, Benefits, Portability and Configuration.

7.2.2.- My Data

The *My Data* section would allow the user to see a summary profile with the most relevant information of your profile (avatar). It would also provide the interface to manage your data.

Data Avatar

This functionality shows, when accessed, the user's personal profile or avatar, with the socio-demographic data, along with interests and preferences based on the data stored in the Data Storage.

The profile is built based on the data provided by the user himself and those that the system infers from other data sources.

Figure 16: Menu mock-up

Data Management

It allows to visualize and edit the personal data that each user provides to the system, either directly or through other applications or modules of EasyPIMS

The data will be displayed on two levels: a first level with the groups of data and a second level that allows access to each data individual.

The data obtained through other modules will inform how and who obtains them and to whom they are delivered.

7.2.3.- Consent Management.

This function allows you to configure and manage the consents to requests coming from external sources to use your data.

Consent can be required for a set of unitary data (e.g. date of birth, marital status, income level) or for a set of them (e.g. socio-demographic data) according to the groups defined in PIMCITY.

Each time that a consent is required from the user, the system will check that the applicant's data are present (Name, Type of Organisation, Website, Country, Accreditations). Moreover, it will inform about the purpose of the processing, whether these data are going to be transferred to third parties, the duration of the consent, where the data are going to be stored and the price or benefit derived from the transfer. Finally, it will provide a link to the detailed conditions where it must be explained how to exercise the rights.

The system that manages the consents will label each request received with a visual code (color, stars...) according to the degree of risk that is estimated for each request. This will allow the user to have visual information to help him/her make a decision with a higher level of information before accepting.

7.2.4.- Benefits and activity record

This function is in charge of keeping the historical record of the data consents accepted by the user. For each record, it will be shown: the date, the organization's data to whom it has been given, the price and all the information associated with the consent and to whom it is requested.

When this function is activated, a summary will be shown that synthesizes and quantifies this information and will be displayed in two levels, one in table mode with one consent per row and another one that will show all the data of a specific consent.

Active consents will be marked with a different color or background than consents that are no longer active.

7.2.5.- Data Portability

This function will allow the user to add/remove new data to/from the system. The data can be imported in a timely manner through a file or from applications that may be connected to the system through the PDK (Internet browsing data, activity monitoring, mobility, etc.).

An important aspect is to normalize how different types of data are recorded or imported, in particular records of mobility, purchases, financial data, etc., so that, data from different sources, can be imported consistently.

7.2.6.- Configuration

The UD will allow to configure at least these parameters:

Access Control

The access control must be simple but safe because it is through the UD where we will manage our privacy (what data we give, to whom and for what). It will allow you to change the passwords, username and other access controls.

Alerts

This functionality allows us to configure how we want the external requests to reach us to indicate among others these options:

- ✓ Degree of immediacy: if we want them to reach us as they occur or receive only one notice per day with all the requests that have occurred.
- ✓ Channel or channels for the alerts: on-screen warnings, email, SMS ...).
- ✓ Filters: establish filters to classify the alerts by price, purpose, degree of confidence.

Account

In this section you can configure the data that identify the users in EasyPIMS: Name and surname, organization, email, mobile phone, account for payments...

7.2.7.- Transparency Tags

Transparency Tags (TT) are used both to label applications, websites or services that we access and to label those who request consent.

Figure 17: Transparency tags mock-up

The label has two parts: the upper one serves as corporate identification while the lower one reports a confidence level that moves in a set of stars and a background that changes according to the confidence level from red (cero, one star), orange (2 and 3) to green (4 and 5 stars).

Transparency labels simply encode the confidence level according to the criteria of transparency, information and control.

To obtain more detailed information, the user may just click on the label to see the supplier's details, the global rating on privacy and transparency. In particular, the user will be informed of all the data and its processing required by the GDPR (purpose, transfer of data to third parties, storage of data, legitimacy, etc.).

7.3.- Telco requirements

Due to their role and position in the Internet, telecommunications companies have the technical means both for collecting personal or anonymous user behavioral data, for example by monitoring network traffic, and for letting users enact their choices about data sharing, for example by enabling them to allow or disallow a certain tracking service that monitors their behavior.

Telecommunications providers see potential for EasyPIMS both in business-to-consumer (B2C) and in business-to-business (B2B) scenarios. At the moment, four main opportunities have been identified:

1. B2B: providing business customers with tools for visibility into the data their users (e.g., employees) disseminate when using the web. This is probably valuable for organizations willing to maintain confidentiality over certain parts of their operations like research and development, procurement, participation in tenders, or strategic moves like mergers and acquisitions.
    - This would be a value-added service with a subscription fee.
2. B2C: providing retail customers with visibility into the data they disseminate while using the web, and

- This would be a value-added services to be provided for a recurring subscription fee.

3. B2B and B2C: intermediation. Acting as alternative data collectors, which EasyPIMS would make more attractive for citizens, thanks to the added transparency, ability to effectively manage consent, and potentially provide rewards to individuals in exchange for personal data.

    - Telecommunications providers would make money by selling personal data to the usual parties that habitually purchase it from trackers, for a subscription fee or one-off.

4. B2B and B2C: enrichment. Acting as alternative or complementary data collector, by supplying personal or aggregated user behaviour data to advertisement providers, research institutions, market analysis firms, and in general to organizations that might need such data, but for some reason cannot or do not want to collect it directly.

    - This would be made available through subscriptions for a recurring fee, or one-off for a one-time fee, determined by the characteristics of the data.

7.3.1.- EasyPIMS in business scenario

Businesses and organizations often need to maintain confidentiality over some of their operations (e.g. product research and development, market launches, merger and acquisition plans), and they are willing to invest in technology to achieve that confidentiality.

Most of the resources they invest to that end are spent either on encryption or on cyber security: encryption is employed for preventing unauthorized parties from accessing the contents of communications and data, while cyber security products are typically used for preventing unwanted software from running on corporate devices, or for preventing documents and data from leaving corporate networks.

However, organizations also “leak” information about the topics they are working on, the parties they are interacting with, and the products, or technologies, or materials they are investigating simply through their users’ (e.g., employees and contractors) web surfing patterns and mobile app usage.

These data can be collected by trackers which could then provide them to competitors or other parties, potentially causing damages like the loss of competitive advantage.

Telecommunications companies therefore see an opportunity to fill in the market gap in confidentiality-protecting services. The opportunity exists for two main reasons: first, telecommunications companies have the technical means to detect interactions with trackers; second, their customers already trust them as providers of security solutions and services. Moreover, a tracker-detecting technology could be integrated with cyber-security solutions, for example a SIEM (security information and event management system) to provide visibility, or a web filter to block traffic towards trackers.

For several reasons, businesses are not interested in tools which would allow them to monetize on the data exposed by their employees. First, the organization is not authorized to speculate on personal data of its employees. Second, the organization is rather interested in preventing its business data to leak outside its perimeter. Third, a granular control over

personal data of all employees would require an amount of effort hardly affordable for the organization. For these reasons, business customers would be interested in parts of a PIMS solution, instead of the whole platform. In particular, they welcome tools which allow them to acquire awareness and control on business data flows leaking outside their perimeter in a centralized manner. In EasyPIMS' specific case, these tools are the Dashboard, the Transparency Tags and the Consent Manager,

Through the Dashboard, the new service would allow the organization to leverage Transparency Tags to first understand which services employees contact when they browse the web and quantify the privacy and security risks connected to such services. Second, the organization will be able to customize its Consent Manager to choose which data can be exchanged, and eventually block web services whose connected risks are too high, similarly to what the customers already do about unwanted or dangerous websites.

This protection will have to be done on a per-customer basis, and both automated and manual service models can be explored.

This service would provide customers with visibility over contacted services whose Transparency Tags report high privacy and security risks, and optionally facilitate the blocking of trackers and/or allow the logging of tracking events.

This service therefore will need to provide these functionalities:

1. A web-accessible corporate dashboard with visibility over the trackers detected and potentially the single tracking events;
2. Machine-readable lists of trackers that can be fed into web filtering solutions or firewalls or other network devices (e.g. Flowspec/RFC 5575, RTBH/Remote-triggered black-holing/RFC 5635);
3. Machine-readable lists of trackers and tracking events that can be fed to monitoring systems like SIEMs (e.g. STIX/TAXII, OpenIOC). This last feature will probable be interesting for large and medium-sized organizations, but not for smaller ones.

Smaller organizations need simple tools that require little to no maintenance effort, and typically do not spend time monitoring or analyzing non-production issues. Therefore, smaller organizations need a product that once set up will run automatically without human intervention. Set up costs also need to be minimized, so the product will need to have pre-defined tracker categories and configurations, and out-of-the-box integration with web filtering / firewall products. For smaller organizations, features 1. and 2. above are enough.

Some medium-sized organizations have more sophisticated confidentiality needs. They will require pre-defined tracker categories and configurations, but they will also want to be able to customize them. For most medium-sized organizations, features 1. and 2. above are enough, but some will also require feature 3. (logging & monitoring).

Large organizations also need to automate cyber security and confidentiality as much as possible, but they typically complement automated solutions with dedicated personnel (or specialized third parties) that can react to unforeseen situations, detect events that escape the tools, quickly understand novel phenomena and override erroneous automated decisions.

Furthermore, larger organizations usually want visibility over events and automated rules before they are enforced, because they often have articulate control structures, complex relations with third parties, and heavier regulations.

Large organizations will therefore need features 1., 2. and 3. above.

For all of the business market segments, EasyPIMS will be a value-added service to be sold with a one-time setup fee plus a recurring subscription fee. The recurring fee will be based on the number of users/seats/devices.

7.3.2.- EasyPIMS in retail user scenario

Telecommunications providers also need to explore the possibility of providing retail users (residential/fixed line and mobile) with forms of visibility over the personal information they disseminate while using the web and mobile apps.

Some retail users are already accustomed to similar services: for example, parental control systems that enable parents to see how much time their children spend using their phone, or specific apps, or websites.

Tracking can be detected:

a) at the network level, e.g. by monitoring IP traffic or DNS queries, or

b) at the device level, e.g. by browser plug-ins or device-based traffic interception systems.

EasyPIMS will need to provide either a way to extract tracking data from the existing traffic monitoring systems within the telecommunications providers (a), or a set of device-level tracking-detection components (b).

EasyPIMS will also need to provide end users with a simple and effective user interface to monitor their web browsing habits and the relative tracking, and to exercise their consent choices.

If EasyPIMS allows end users to choose to have the telecommunications providers act as data collectors instead of the usual trackers, then the telecommunications providers will need a way to quantify the “value” provided by each customer’s data. This will enable the telecommunications providers to use that information to reward each customer, e.g. through their loyalty programs.

7.3.3.- Intermediation

While allowing both retail and enterprise customers to choose whether to be tracked, and by whom, telecommunication providers also need to explore the possibility of acting as a new, more transparent data collector with which the users might want to share data.

Technically, this means telecommunication providers will need to provide their users with:

1. a way to enable the telecommunication provider’s trackers and potentially the relative third parties, while optionally blocking other trackers based on users’s choice
2. a way to clearly see what data the users are allowing the telecommunications providers to collect, as well as the parties with which this data might be shared
3. a historical view of the data shared

4. an interface through which to exercise their rights (e.g., data erasure)

Point 4 might be a challenge. While the telecommunications provider will be able to comply with the users' requests by acting on the data it manages directly, some of the data collected by the telecommunications provider might be outside its control. For example, if some data has been collected by the telecommunications provider and then shared with third parties, these third parties will probably have a copy of the data. In this case, the telecommunications provider cannot directly update, modify or destroy that copy of the data, and it will need a mechanism to ask the third parties to update / modify / destroy the data on its behalf.

Citizens would have two main reasons to choose their telecommunications provider as tracker with EasyPIMS:

1. in this scenario the telecommunications providers would be a more transparent tracker that allows users to make effective and meaningful choices over their personal data
2. sharing of personal data could be rewarded with discounts, accessory services or loyalty programs through the telecommunications providers, if economically viable (this will need to be explored)

In this scenario, telecommunications providers would make money by selling the data to advertisement providers and other parties that might be interested in it, either for a recurring fee, or one-off.

7.3.4.- Data collection and enrichment

Telecommunications providers have – or can collect – both personal and anonymous data that might be useful to several parties.

For example, one advertisement provider might want to investigate user behaviour data on a website whose ads are run by a competitor. The competitor would certainly not want to share these data with the advertisement provider, but telecommunications providers might be able and willing to do it.

In other markets, a technical research institution might be interested in the performance of certain network protocols, or a humanities institution might be interested in aggregated network behavior patterns that could enable sociological analyses by geographical region or other characteristics.

Technically, telecommunications providers could produce relevant data to this end in two ways:

a) by collecting data on individual users or devices, or

b) by collecting data which is aggregated (representing not one but several users, not personally identifiable)

Enrichment with personal data

The extent to which telecommunications providers are allowed to collect data on individual user behaviour will need to be determined during the project, as well as any requirements that may arise regarding and user notice / consent and data protection.

These data might for example include the date and time a user lands on a website, the subsequent connections / websites / trackers the user's device contacts, the duration of the user's interaction with the website, and an indication of what the user does after interacting with the website (e.g. open up an affiliate, or competitor, website / stop browsing / etc.).

These data might be collected at the network level, e.g. through IP traffic capture, thereby reusing equipment the telecommunications providers already have, or by developing software components to be installed on the users' devices (this would have to be done by other parties in the Consortium, or by third parties once the project ends).

Enrichment with non-personal data

Non-personal data poses fewer privacy-related challenges, at least from a regulatory requirement standpoint.

These data might for example include statistics about the interactions of a certain user demographic, or the whole customer base of the telecommunications provider, with a certain website. The data might show the amount of visitors, or of traffic, that a website receives during the course of the day, or of the week etc., or on the occasion of particular events (a football match, a weather phenomenon, a political event etc.), and/or the amount of time the average user (or the average user belonging to a certain demographic). The data might also include statistics about what the users – on average – do online *before* opening the website, during the interaction with the website, and after closing the website.

Such data might be useful to parties interested in how and when a website is attracting, or failing to attract, user traffic (or traffic from a certain demographic). Again, these data would mainly be valuable to parties that do not control or track the website, for example advertisement providers who do *not* have the website as client.

These data might be collected at the network level, e.g. through IP traffic capture or sampling, and aggregated before being shared, so that no personally-identifiable information leaves the telecommunications provider.

Collection of telco-specific data

Telecommunications providers also have other types of data that could be useful to other parties, or for internal marketing purposes.

These data include the geographical position and movements of users as measurable by the network (like cellular network antennas, Wi-Fi antennas, and other types of network). Such data can be anonymized, aggregated, or grouped in clusters by similarity or arbitrary metrics.

Anonymous or otherwise, these data can be used to monitor the flow of people, of vehicles, and of mass transportation along roads and other transportation means, at venues or in correspondence with particular events or conditions (meetings, concerts, weather events). Based on these data, patterns can be discovered in seasonality and in the response to events or conditions; trends can be monitored and forecasts can be made.

Transportation services, shopping centres, the tourism industry, venues, museums, vacancy resorts, leisure centres, the retail industry, banks and insurance companies, advertisement providers, as well as academia, could all benefit from these data to better

model their customer audiences and make more informed marketing or service management decisions, or to achieve better understanding of social dynamics.

Telecommunications providers could use these data for internal marketing purposes, sell it for a fee, or pursue shared-revenue models with interested parties.

7.3.5.- Marketplace

In the intermediation, enrichment, and collection scenarios, telecommunication providers will need a way to learn about the demands of potential buyers of personal, aggregated, or anonymous data, and a marketplace to make deals and supply the data. This could be accomplished by developing and running a data trading platform.

Given the volumes of data involved and the speed at which supply, and demand might change, the trading platform will need to be fully automated and provide remote APIs to enable all parties to place bids and buy and sell data programmatically.

7.4.- Technical requirements

The EasyPIMS platform will be designed, implemented and tested to demonstrate the flexibility and usability of the components developed in the PDK. However, it presents its own technical requirements and limitations that will be taken into account since the beginning of the project. Moreover, EasyPIMS will not only use the different PDK modules but it will also develop new tools from scratch.

Similar to the technical requirements for the PDK, the different partners of the consortium will offer their expertise in the different technological fields needed to develop the final integrated solution. NEC, FW and TID has expertise in capturing and analyzing network data, LSTECH and ERMES in cloud integration and CLIQZ and GDATA in data trading and privacy preserving data commercialization.

The methodology followed to gather the technical requirements mimic the one followed for the PDK. Since the EasyPIMS platform is based on the elements of the PDK, that will be designed by month 9, in this document, we only present a preliminary version of those requirements, that are based on those of the PDK tools, and they will be updated in the deliverable D1.2 (month 18).

7.4.1.- Goals and objectives

The main goal of the EasyPIMS is to design and develop a platform to allow users of telco providers to trade their data with third parties in a controlled way. To this end it will:

- ✓ Provide the technical platform which will both run EasyPIMS and make it easy for future components to be plugged into the system, thereby creating a fertile ground for experimentation from third parties.
- ✓ Design, document and implement open APIs and SDKs to ease integration between modules, and with future components from third parties.
- ✓ Design and develop an orchestrator that allows the execution and interoperability of the different building blocks in existing computing platforms like AWS

- ✓ Design and implement the EasyPIMS components (TT, PDA, Marketplace) and integrate them.

Moreover, a complete set of user friendly dashboards will be designed to allow the interaction of the users.

7.4.2.- Proposed tools

Personal Data Avatar (PDA)

It is a digital projection of data stored in the P-DS, under full control of the user. PDA contains a set of information synthesized by ad hoc analytics fed with the data made available by the user in the P-DS and the settings specified in the CM. In other words, the PDA is a user-controlled privacy-preserving P-DS synthesis. The PDA is the interface between the user and the services. Thanks to the PDA, the user becomes the only owner of her data and acquires the freedom and power of deciding which data to share with which service. The PDA lets the user know the actual monetary value of exposed data (obtained through the D-VT component), increasing consciousness regarding her privacy. In turn, the PDA provides services (e.g., advertisers) with personal information built and validated by the user, thus more precise than those extracted nowadays by opaque tracking companies. Thanks to this, services considerably reduce their costs: they reduce expenses for obtaining users' data from trackers and data brokers. And they increase their reputation as trusted parties by eliminating the risk of violating privacy.

Transparency Tags (TT)

It is the analogue to a Nutrition Label for food which provides the information about the ingredients, their provenience, intolerance risks, etc. of food. The TT communicates in an easy to understand way the nature of the web service the user is accessing to. For each web service (e.g., a website, a web service, a mobile app) EasyPIMS exposes the information computed by the P-PM, such as its owner, its purpose, the personal data it collects, etc. This is summarized in scores - automatically computed by the P-PM specialized analytics - revealing the potential privacy risk associated to that website.

User Dashboards (UD)

Users need simple and intuitive means to access and control their information. For this, EasyPIMS implements a digital dashboard to keep control of their personal data in the P-DR, check their value via the PDA, controlling service reputation via the TT, etc. Thanks to the UD, users are able to control the EasyPIMS, their data, and their preferences. The UD will be designed following state of art approaches, with simplicity in mind. Accessible via an interactive dashboard, specifically designed for smartphone usage, the UD will be the revolutionary means to take control of personal information on the web. Modular by design, private by design, it builds upon the TT and the PDA to increase awareness in users.

Marketplace

To interoperate with services, EasyPIMS implements a simple digital Marketplace, i..e, the EasyPIMS Data Broker. Similar to a physical Marketplace, the EasyPIMS Marketplace facilitates the trading of user data (via the PDAs) and web services. EasyPIMS collects and offers in the Marketplace profiles obtained by PDAs made available by users. Web services provide the capability to check and purchase profiles (via the D-TE component). This allows conducting transactions in a transparent way andproviding revenues to users.

Cloud Controller

The cloud controller is the platform orchestrator where EasyPIMS will be executed. On the one hand, this component integrates the different modules of EasyPIMS to make them work together in a commercial Cloud provider. On the other hand, it provides the connection with the internal data silos available inside the Telco provider running EasyPIMS, ensuring the platform only can access the information for which the user has previously provide her consent.

Open API

Integration among components is key for the distributed systems to work properly. EasyPIMS will implement the APIs of the SDK designed to seamless integrate components and enable their interoperability and re-usability. Open APIs will be designed putting particular attention on three key features: privacy, security and scalability. To achieve this, we first classify components based on their purpose (e.g., analytics, UI, data extractors, etc.), and we then define per-class requirements and API prototypes as well as their integration SDK. Then, each partner responsible for the design of a component will be in charge of designing and implementing its corresponding APIs.

7.4.3.- Functional requirements

Requirement Name	Mandatory / Obligation / Recom-mended	Description	Elements Involved	Notes
R1: Login	M	Users shall be able to login in the system.	Cloud Controller, UD	Actor: Users, Data Buyers
R2: User Menu	M	The system will provide a menu with the possible sections available to the user	UD	Actors: Users
R3: Insert/Modify data sharing preferences	M	Users shall be able to give or remove consent to the sharing of content. It can be done automatically by	UD	Actor: Users, others system elements

		other modules (e.g., P-CM)		
R4: Consult the data users provided to the system	O	The users shall be able to consult the data stored into the system in an easy to understand way	P-DS	Actor: Users
R5: User profile	O	The platform will provide a summary of the user with information of their profile	Cloud Controller, PDA	
R6: List Active Consents	O	Users shall be able to see all the consents active on their data. By default, unless a consent is created all data is not subjected to trading.	UD, PDA	Actor: Users, Other systems
R7: Create New Consent	O	Users shall be able to create a consent for a data element or type, with different constrains such as: time-to-live, price, purpose, etc.	UD	Actor: Users
R8: Remove Consent	O	Users shall be able to remove constrains.	UD	Actor: Users
R9: Data Value accessibility through a web interface	R	To grant access to the data in the DVTMP module to not skilled users, it should implement an	UD	

		intuitive web interface for this purpose.		
R10: Data Buyer Authentication	M	The Data Buyer authenticates on the TE to start operating	Marketplace	
R11: Data Buyer Menu	M	The platform will provide a menu with the options available to the data buyers.	Marketplace	Actors: Data buyers
R12: Data Buyer exploration	M	The Data Buyer queries the TE to test if there are sufficient Data Sellers to buy data from for the desired data type and for the given segmentation parameters	Marketplace	
R13: Create an Offer	M	The Data Buyer creates a Draft Offer on the TE to buy Sellers' Data specifying institution, purpose, price, the requested data type and segmentation data as well as the budget or desired data amount	Marketplace	
R14: Publish an offer	M	The Data Buyer publishes the Draft Offer so that the TE identifies, using R2: Explore, all matching Data Sellers and defines the strategy to transact with each one	Marketplace	

		according to his D-CM configuration	
R15: Notify Data Sellers	M	Triggered by the strategy for Data Sellers that require explicit consent, the TE notifies all matching Data Sellers about the recent published Offer	Marketplace
R16: Receive notification	M	The Data Seller receives the notification regarding the intention of the Data Buyer to buy his data	Marketplace
R17: Accept an offer	M	The Data Seller indicates that they want to sell the required piece of data to the Data Buyer notifying the TE about it	Marketplace
R18: Reject an offer	M	The Data Seller rejects the Offer notifying the TE about it	Marketplace
R19: Transact with consent	M	Triggered by the response in R17: Accept an offer, the TE requests the pieces of information to the Data Seller's DPA and makes it available to the Data Buyer while also ensuring that the Data Seller receives the according value.	Marketplace, DPA
R20: Finish Transaction	M	Triggered by the strategy for Data	Marketplace, DPA

		Sellers that don't require explicit consent, the TE request the pieces of information to the Data Seller's DPA and makes it available to the Data Buyer while also ensuring that the Data Seller receives the respective value.	
R21: Discard matching Data Seller	M	Triggered by the response in R18: Reject offer, the TE discards the matching Data Seller that rejected the Offer.	Marketplace
R22: Transaction history	M	The Data Seller accesses the transaction history being able to visualize all Offer parameters and the status of each transaction.	Marketplace
R23: Offer progress	M	The Data Broker accesses to the current state of an Offer being able to visualize how many transactions happened so far.	Marketplace
R24: Unpublish offer	M	The Data Broker unpublishes an Offer when he decides to do so, making it unavailable to Data Sellers	Marketplace

R25: Unpublish offer by condition	M	The TE decides to unpublish an Offer after a budget goal or amount of data sets is reached, making it unavailable to Data Sellers	Marketplace	
R26: Support standard data interchange formats	R	The platform should support the standard data interchange file formats for contact information, like VCF	Open API	Standards will be considered and we will support the most common ones. A mechanism that will allow the incorporation of new standards will also be considered.
R27: Support basic and generic data types	M	The metamodel shall be able to support all the major data types that might be used in the various sources.	Open API	The metamodel shall accept numbers (various types), strings (in general), strings of specific formats (such as urls, emails, etc), json, arrays etc.
R28: Data transformation and mapping between the sources and the platform	M	A metamodel for the possible transformation of the source data to the allowed/ supported by the PIMCITY platform shall be defined. Specific mappings from the source data and the PIMCITY data model shall be defined.	Open API	Also, the definition of the output types that will be needed when

R29: Support Data characterization	R	The metamodel should allow the characterization of the data attributes in order to support the data aggregation and the privacy functions	Open API	For example, we should be able to characterize data with various categories (to be defined) like “sensitive” or “aggregated” or “non-analyzed” etc. This will allow for better filtering, selection and control of the processes/ functions that can be applied to the data, the sharing processes etc.
R30: Data import flow	M	A mechanism that will define the flow of the data import process	Open API	All the steps that need to be followed in order for the data to be stored in the platform, including transformations, mappings, aggregations (If needed)
R31: Data privacy levels	R	The import mechanism should support a number of privacy levels that will define the detail of the data to be transferred or the aggregation level etc.	UD	This will be used to define the available actions per dataset and the aggregation/ sharing functions
R32: Data aggregation functions	M	A set of data aggregation functions shall be defined for data	Cloud Controller	Functions like sum, avg, count,

		types that allow aggregations		categorization etc
R33: Data anonymization functions	M	A set of data anonymization functions shall be defined for specific data types	Cloud Controller	
R34: Data import/ export APIs	M	Provide APIs to allow easy and secure data import and export/ sharing	Open API	
R35: Automatic User profiling	O	The platform should be able to automatically generate the profiles	PDA	It will be based on navigation data
R36: Query per-service data in the system	M	System elements shall be able to query data about web services	Transparency Tags, Open APIs	Actor: Other system elements
R37: Update per-service data in the system	M	System elements shall be able to modify data about web services	Transparency Tags, Open APIs	Actor: Other system elements
R38: Provide revenues to users	M	Marketplace shall reward the users any time their PDAs are queried	Marketplace, UD, Open APIs	Actor: Users
R39: Consult Transparency Tags	M	User dashboard shall provide users the capability to consult Transparency Tags	UD, Transparency Tags, Open APIs	Actor: Users
R40: Consult Marketplace Transactions	M	User Dashboard shall provide users the capability to consult purchase transactions between PDA and services	UD, Marketplace, PDA, Open APIs	Actor: Users, Data buyers

7.4.4.- Non functional requirements

As in the case of the technical requirements, the non-functional requirements are based on the PDK ones However, theypresent several differences. Since the EasyPIMS is designed to be used by end users and the PDK is designed to be used by companies, there are several differences.

NFR1: Easy to use and understand. The User Dashboards should provide easy to use and easy to understand interfaces.

NFR2: User Friendly. The different parts that interact with the end users should present User Friendly UI, clean and simple layouts. It is also important that all the information presented to the user (including the privacy policy or the consent management) should be written in a way that normal users can understand.

NFR3: Scalability. The system should be able to handle at least 100.000 users (not using the system at the same time). Moreover, it should support at least 1000 different data trading operations per second.

NFR4: Time constraints. The different modules should provide results in a reasonable time that allow the execution of different applications presenting time constraints.

NFR5: Legal compliance. The whole data processing (from the consent to the data trading) should comply with the GDPR and the rest related European laws.

NFR6: Portability. The system should provide the possibility to be controlled in different devices such as mobile devices, tablets or laptops.

NFR7: Cloud. The system should work in standard cloud providers.

NFR8: Data security. The data should be stored in an encrypted format and following extract data protection policies.

7.4.5.- Constraints

The design of the software modules for EasyPIMS will respect several constraints to ensure a robust, yet flexible architecture, high-quality code and secure operations.

Modules will interact using web-based Open APIs. As such, we must ensure that all communications are private and secure, and methods for authenticating users are robust. For this, we will leverage state-of-the-art best practices to design Open APIs. All communications will be encrypted using HTTPS/TLS1.3 and we will make use of authentication tokens (OAuth) to enforce control on the use of the APIs.

Some modules (e.g. Transparency Tags, Data Marketplace) will operate with large amounts of data, potentially from millions of users and thousands of data buyers. Hence, these modules must scale well and must be designed to be part of a distributed architecture, so that they can handle small to large communities of users with minimal effort. To this end, we will use scalable-by-default tools, such as Django, MongoDB, Spark.

Finally, all the phases of development must follow the best practices for good quality development. We will follow the agile software development approach and the DevOps

paradigm for automated build, test, continuous integration, and continuous delivery. All the code will be open source, and available on the popular repository (e.g., GitHub and GitLab).

7.4.6.- High level architecture

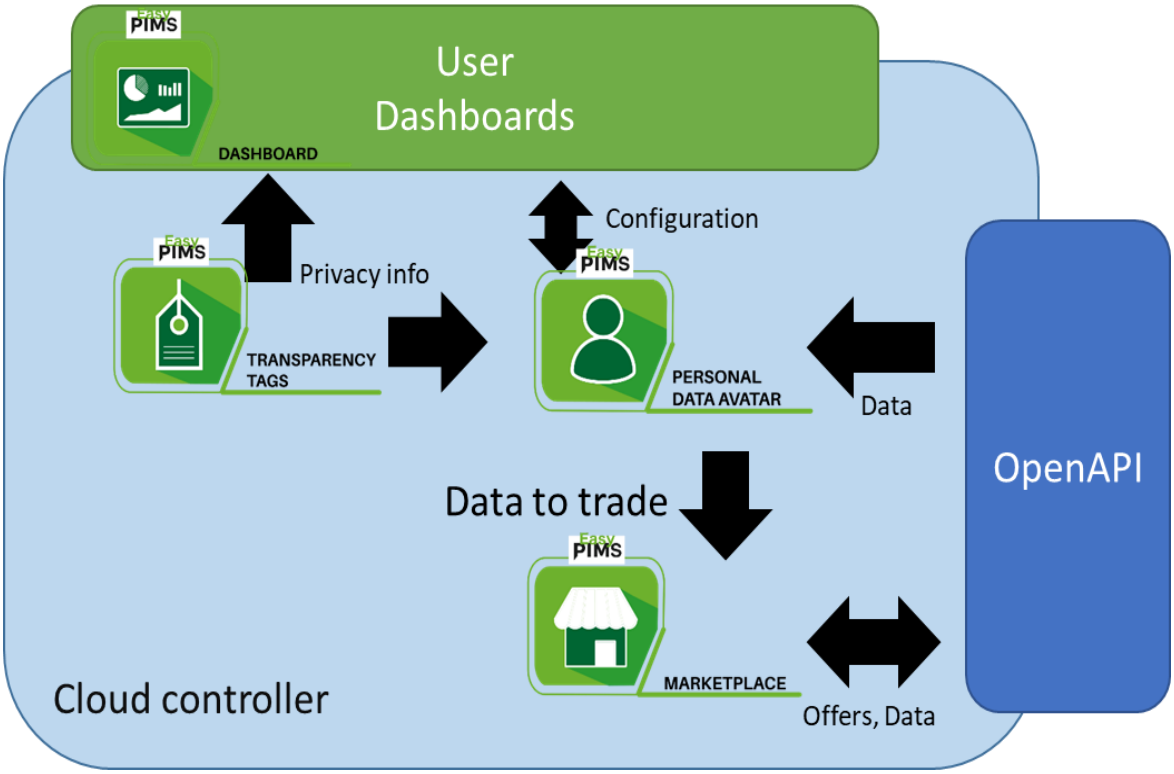

Figure 18: High level architecture of EasyPIMS

7.4.7.- Interfaces

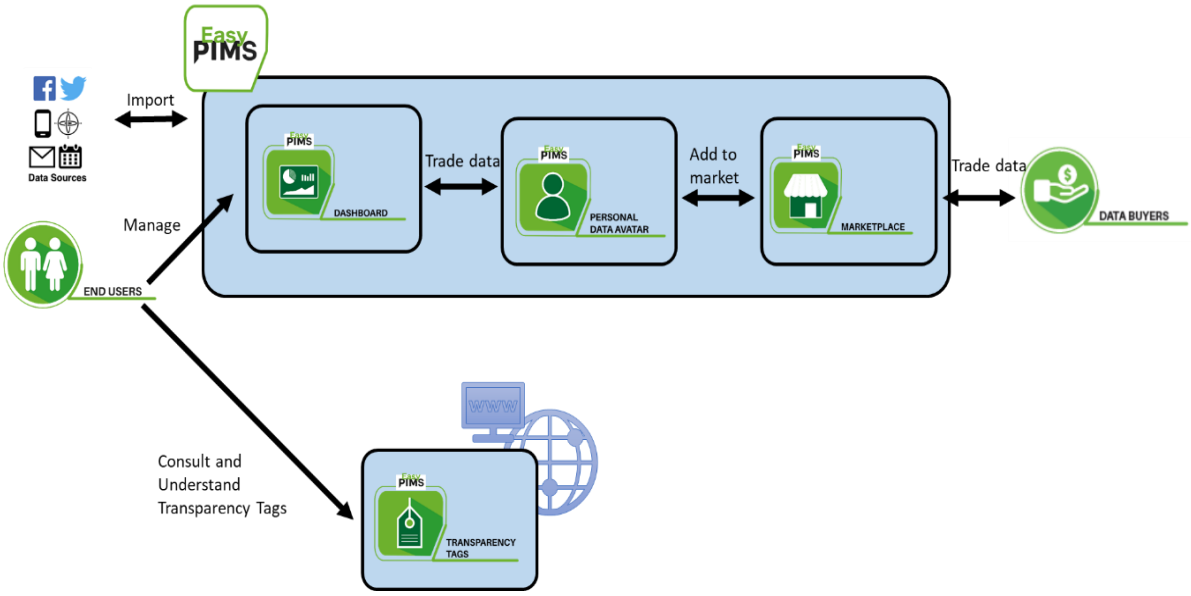

Figure 19: Interfaces of EasyPIMS

8.- CONCLUSION

This document entails the requirements that will form the basis of the PIMCity Project. First, it presented the Final Users requirements. The consortium considers that meeting the needs of end-users is key for the success of the Project. To this end, a focus group with real users was consulted about their interests in PIMS systems. Moreover, the document elaborated on the GDPR, e-Privacy Directive and all the relevant rules for the processing and trading of personal data. Then, it presented the main needs of the current PIMS to adopt some of the software pieces developed within the project. Furthermore, given the user centered model proposed by the users during the focus group, the document also analyzed the requirements for a user centered data economy. Finally, the requirements from the different stakeholders were translated into technical requirements creating the specification of different tools that will be developed by the technical Work Packages.

Moreover, as for the case of the PDK, this document presented a preliminary version of the requirements for the EasyPIMS platform, including again the opinion of the final users and the specific needs of the telco providers to can run such a platform. These requirements will be updated in Deliverable 1.2 after the design of PDK components provide more insights of the limitations a platform like EasyPIMS may have.

The consortium will work towards meeting the identified requirements. Meeting the requirements will contribute to achieving the overarching objectives of PIMCity, namely: to provide tools for the development of new PIMS and to demonstrate its usability with the creation of EasyPIMS. This will be to the benefit of both citizens (as they can participate from the data economy) and PIMS companies (as they can improve the technology of their systems.). Furthermore, the inclusion in this document of a summary of the legal requirements (completed in D7.1) applicable to the PDK and EasyPIMS aims to ensure that companies using the components will also comply with the GDPR and the rest of European laws. Partners will constantly monitor and adjust their course of action where needed to guarantee compliance with the law.

As the work for developing the PDK components evolves, requirements defined in this document may be subject to amendments. Deliverable 1.2 will be used to assess the requirements defined in this document and, based on the findings of the assessment; requirements may be supplemented, amended or deleted.

Appendix A: End Users Focus Groups.

OBJECTIVE

The aim of the two working groups set up by AUI is to have the views of end users on their perception of privacy and on the PIMCity project in those aspects that affect it.

METHODOLOGY AND SCHEDULE

The people who have participated have done so voluntarily without receiving any economic compensation and have been chosen from among the associates resident in Madrid so that in each of the groups people from each generational age group and that there is no more than a 20% difference between people of the same gender. E

Four meetings were held with them, two face-to-face (20th and 27th February) and two virtual (12th and 19th March):

- ✓ 20/02: Presentation of the project and issues to know the perception of privacy were addressed in this first meeting.
- ✓ 27/02: The second meeting focuses on regulatory knowledge (RGPD), the perception of the value of their data and the identification of these data.
- ✓ The third meeting focuses on analyzing different labeling systems and see what considerations can contribute to the PIMCity modules that in one way or another should be used directly by users (Data Avatar, Transparency Tags and EasyPIMS)
- ✓ The last one is dedicated to sharing the results and conclusions of the work done.

Appendix B: PIMCity Questionnaire about "Privacy Perception".

This Questionnaire conducted by AUI among its associates on april 2020 from which to gain insight into the following questions:

- What are consumers' overall sentiments toward data privacy? What are their main concerns?
- How much do consumers understand about how data (particularly location data) is used and collected?
- What does personalization mean to consumers, and are they willing to share data for increased personalization of products and services?
- What, if anything, would encourage consumers to be more willing to share data, and which types of data are they most willing to share?

To answer these questions, we send the questionnaire to 1200 users users between the ages of 18-65 in Spain and we get 243 responses.

The questions that were posed and their possible answers are listed below.

Q1. Are you concerned about data privacy?

Younger people, Generation Z (1995 - 2015) and Millennials (1980-1994) are less concerned and pay less attention to privacy, are more willing to give up and share their personal data and are more likely to benefit from it.

Are you concerned about data privacy?					
	Gen Z (1995 - 2015)	Millennials (1980-1994)	Gen X (1970 -1979)	Baby Boomers (1957- 1969)	Total
Not Concerned at all	9%	9%	6%	2%	6
Slightly concerned	20%	25%	20%	16%	20
Neutral	18%	16%	12%	8%	14
Somewhat concerned	25%	22%	27%	26%	25
Very concerned	28%	29%	35%	47%	35

Q2. If you have concerns about data privacy, what are they?

Digging into what constitutes concerns about data privacy, consumers listed identity theft/fraud and stolen passwords most often. Generational differences aren't a factor here – the top two concerns for every generation were fraud and stolen information.

If you have concerns about data privacy, what are they?	
Identity theft and fraud, Stolen passwords	72%
Stolen passwords	64%
Not knowing what personal information is being used for	59%
Information being sold for profit	54%

Location tracking	53%
Not knowing how personal information is collected online	52%
Unsolicited marketing communications	41%
None of the above	7%

Q3. What do you consider to be personal information?

Home address and phone number top the list, while online or mobile browsing habits landed at the bottom.

What do you consider to be personal information?	
1.Home address	98%
2. Phone number	82%
3. Birthdate	77%
4. Location data	75%
5. Email	70%
6. Name	66%
7. Messages sent via messaging apps	55%
8. Work address	49%
9. Online or mobile browsing habits	48%
10. None of the above	10%

Q4 How do you think your location data is used ?

The majority of respondents understand location data is used to provide targeted or tailored advertising and offers

Q4 How do you think your location data is used ?	
To provide targeted ads,	62%
To deliver relevant search results,	48%
To help companies better understand their users,	41%
To create personalized product experience	40%
To improve public safely products	23%
None of above, Others	11%

Q5 Are you comfortable sharing my location data for marketing purposes?

Thirty- four percent of respondents feel comfortable sharing location data for marketing purposes, and 27% are indifferent.

Are you comfortable sharing my location data for marketing purposes?	
Strongly agree	10%
Somewhat agree	24%
Neither agree nor disagree	28%
Somewhat disagree	25%
Strongly disagree	17%

Q6 I believe that my data is used to personalize experiences and products

Fifty-nine percent of respondents understand their data is used to personalize experiences and products.

I believe that my data is used to personalize experiences and products	
Strongly agree	21%
Somewhat agree	38%
Neither agree nor disagree	30%
Somewhat disagree	8%
Strongly disagree	4%

Q7 I believe that companies/ brands benefit from data collection and usage.

However, 63% of respondents believe companies and brands benefit from data collection and usage, while only 43% believe that consumers benefit.

I believe that companies/ brands benefit from data collection and usage	
Strongly agree	30%

Somewhat agree	33%
Neither agree nor disagree	26%
Somewhat disagree	6%
Strongly disagree	5%

Q8 I believe that consumers benefit from data collection and usage.

The data shows that the current conversation around data privacy isn't doing justice to the benefits of personalization and the consumer desire for personalized content. most respondents think that personalization benefits companies more than citizens

I believe that consumers benefit from data collection and usage	
Strongly agree	16%
Somewhat agree	28%
Neither agree nor disagree	35%
Somewhat disagree	14%
Strongly disagree	8%

Q9 Where do you prefer to see personalized content?
Despite claiming they're uncomfortable sharing with social sites, consumers want to see personalized content within them, which makes sense, considering they spend a lot of time each day on social media.

Where do you prefer to see personalized content?								
Google	Amazon	Facebook	Mobile Apps	Webs	Browsers	Email	Major Publishers	Other Social
44%	42%	32%	27%	25%	19%	15%	8%	5%

Q10. How willing are you to share data with companies knowing that the link between the data shared and the benefits provided are clear?

Eighty-three percent of respondents are either open or neutral in their opinion about sharing data with companies if the link between the data shared and the benefits provided to them are clear.

How willing are you to share data with companies knowing that the link between the data shared and the benefits provided are clear?	
Strongly agree	13%
Somewhat agree	39%

Neither agree nor disagree	31%
Somewhat disagree	12%
Strongly disagree	6%

References

Royal Statistical Society. (2017). *Data governance: public engagement review.* Retrieved from https://www.thebritishacademy.ac.uk/publications/data-ai-management-use-governance-21st-century/

Cardogan, R. (2004). An Imbalance of Power: The Readability of Internet Privacy Policies. Journal of Business and Economic Research.

DMA-Acxiom. (2018). *GDPR: A consumer perspective.* Retrieved from https://dma.org.uk/research/gdpr-a-consumer-perspective

Doteveryone. (2018). *People, Power and Technology: The 2018 Digital Understanding Report.* Retrieved from Miller, C., Coldicutt, R., & Kitcher, H. for Doteveryone. : https://doteveryone.org.uk/report/digital-understanding/

EC Eurobarometer The General Data Protection Regulation. (2019). *Special Eurobarometer 487a: The General Data Protection Regulation.* Retrieved from https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/survey/getsurveydetail/instruments/special/surveyky/2222

EC Eurobarometer: Online platforms. (2016). *Special Eurobarometer 447: Online platforms.* Retrieved from http://ec.europa.eu/information_society/newsroom/image/document/2016-24/ebs_447_en_16136.pdf

EU Eurobarometer e-Privacy. (n.d.). *Flash Eurobarometer 443: e-Privacy.* Retrieved from https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/FLASH/surveyKy/2124

Harris Interactive. (2019). *Adtech - Market research report.* Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0023/141683/ico-adtech-research.pdf

Illuminas for Citizens Advice. (2016). *Consumer expectations for personal data management in the digital.* Retrieved from https://www.citizensadvice.org.uk/Global/CitizensAdvice/Consumer%20publications/Personal%20data%20consumer%20expectations%20research.docx.pdf

Information Commissioner's Office. (2019). *Information Rights Strategic Plan: Trust and Confidence.* Retrieved from https://ico.org.uk/media/about-the-ico/documents/2615515/ico-trust-and-confidence-report-20190626.pdf

Internet Society - Ipsos. (2019). *The State of user Privacy and Trust Online.* Retrieved from https://www.internetsociety.org/resources/doc/2019/the-state-of-user-privacy-and-trust-online/

Into The Minds. (2018). *Reading privacy policies of the 20 most-used mobile apps takes 6h40.* Retrieved from https://www.intotheminds.com/blog/en/reading-privacy-policies-of-the-20-most-used-mobile-apps-takes-6h40

Ipsos Mori. (2016). *Digital footprints: Consumer concerns about privacy and security.* Retrieved from https://www.communicationsconsumerpanel.org.uk/research-and-reports/digital-footprints

MacDonald, A. a. (2008). Retrieved from The Cost of Reading Privacy Policies.: https://kb.osu.edu/handle/1811/72839

Ofcom. (2019). *Adults Media Use and Attitudes report - data tables.* Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0026/149840/adults-media-use-attitudes-2019-data-tables.pdf

Open Data Institute . (2019). *Attitudes towards data sharing. .* Retrieved from https://pubmed.ncbi.nlm.nih.gov/29653503/

Telefonica - IE . (2020). *Data, Privacy, and the Individual.* Retrieved from https://www.telefonica.com/documents/341171/0/Report+Privacy+matters/2c4d9b13-ba1c-d5b4-9997-86cf1fe5e875

The Centre for Data Ethics and Innovation. (2020). *Review of online targeting: Final report and recommendations.* Retrieved from https://www.gov.uk/government/publications/cdei-review-of-online-targeting/online-targeting-final-report-and-recommendations

Which. (2018). *Consumer research on attitudes to data collection and use.* Retrieved from https://www.which.co.uk/policy/digital/2707/control-alt-or-delete-consumer-research-on-attitudes-to-data-collection-and-use