“Building the Next Generation Personal Data Platforms”
G.A. n. 871370

DELIVERABLE 6.5
PIMCity Training Kit

H2020-EU-2.1.1: PIMCity
Project No. 871370
Start date of project: 01-12-2019
Duration: 33 months

Due Date of Delivery: 31/08/2022
Date of Delivery: 31/08/2022

Document Information

Document Name: **PIMCity Training Kit**
Deliverable Type: RTD
Dissemination Level: Public
WP6 – Title: **PIMCity Training Kit**
Task 6.3
Revision: 1
Revision Date: 31/08/2022
Authors: Miguel Pérez (AUI) as main editor, and all partners involved in WP6.

Lead Partner: AUI

Dissemination Level

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

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

Approvals

	Name	Entity	Date	Visa
Author	**Miguel Pérez**	**AUI**	14/08/2022	
WP Leader	**Miguel Pérez**	**AUI**	14/08/2022	
Reviewer	**Ruben Cuevas**	**UC3M**	29/08/2022	
Coordinator	**Marco Mellia**	**POLITO**	31/08/2022	x

Document history

Revision	Date	Modification
Version 0	15/08/2022	Initial version
Version 1	26/08/2022	Version for review
Version 2	30/08/2022	Reviewed version
Version 3	31/08/2022	Final version

Disclaimer

The user uses the information at its sole risk and liability.

List of abbreviations and acronyms

Abbreviation	Meaning
G.A.	Grant Agreement
CA	Consortium Agreement
GA	General Assembly
PB	Project Board
PC	Project Coordinator
PO	Project Office
IR	Interim Reports
DCP	Dissemination and Communication Plan
PIMS	Personal Information Management Systems
PDK	PIMS Development Kit
P-CM	Personal Consent Manager
P-PM	Personal Privacy Metrics
P-PPA	Personal-Privacy Preserving Analytics
DVTUP	Data Valuation Tool from the Users' Perspective
DVTMP	Data Valuation Tool from the Market Perspective
DTE	Data Trading Engine
DPC	Data Portability and Control
DP	Data Provenance
DKE	Data Knowledge Extraction
DA	Data Aggregation

Document index

List of figures

Executive summary

This deliverable presents the Collection of training materials and resources for the use and exploitation of PIMCity outputs. One of the goals of the project was indeed to reach out to relevant stakeholders outside the project consortium to give them specific training on the PIMCity methodology and tools. Training activities were structured to satisfy the needs and interests of technical stakeholders (programmers, engineers, professionals etc) and decision makers.

This report presents the training material prepared for Task 6.2 "Training and tutorials" of the PIMCity project. This deliverable will contribute to maximize opportunities of adoption, increasing public acceptance and build new awareness and educational opportunities around personal data platforms.

The training materials will address three different target groups: Developers who want to use or integrate one (or several) of the developed modules, trainers who will train others in the use of the developed technologies and institutions interested in the implementation of data exploitation systems committed to the rights of users.

For **companies, entrepreneurs, developers, integrators, and other stakeholders** interested in the digital and data user-centric businesses, we deliver online video tutorials providing guidance for the use of the tools developed on the PDK, as well as demonstrating its advantages and benefit to end users and other stakeholders. These video tutorials and materials are accessible through **PIMCity Youtube channel** as well as in the **project website**, and other divulging sites associated to the members of the consortium.

For society at large, we have produced educational materials and sessions for engaging citizens into project topics (e.g., privacy, data ownership, etc.) in a broader scope than the innovation carried out in the project

Finally, universities (UC3M, KUL amd POLITO), will deliver training material that will become part of their master's courses and degrees (e.g., Big Data, Cybersecurity, Data Science).

The document goes through the different modules that form the PIMS Development Kit (PDK): Tools to improve Users Privacy, tools for the new Data Economy and tools for Improved Data Management. At the end are included links to videos, articles, brochures, and presentations that help to have a more general vision of the project.

PIMS Development Kit (PDK) components

This section describes each of the PDK components grouped into three blocks: Tools to improve Users Privacy, tools for the new Data Economy and tools for Improved Data Management

From this URL: https://easypims.pimcity-h2020.eu/ one can access to test, watch the videos, consult the APIs or download the software of each PDK component.

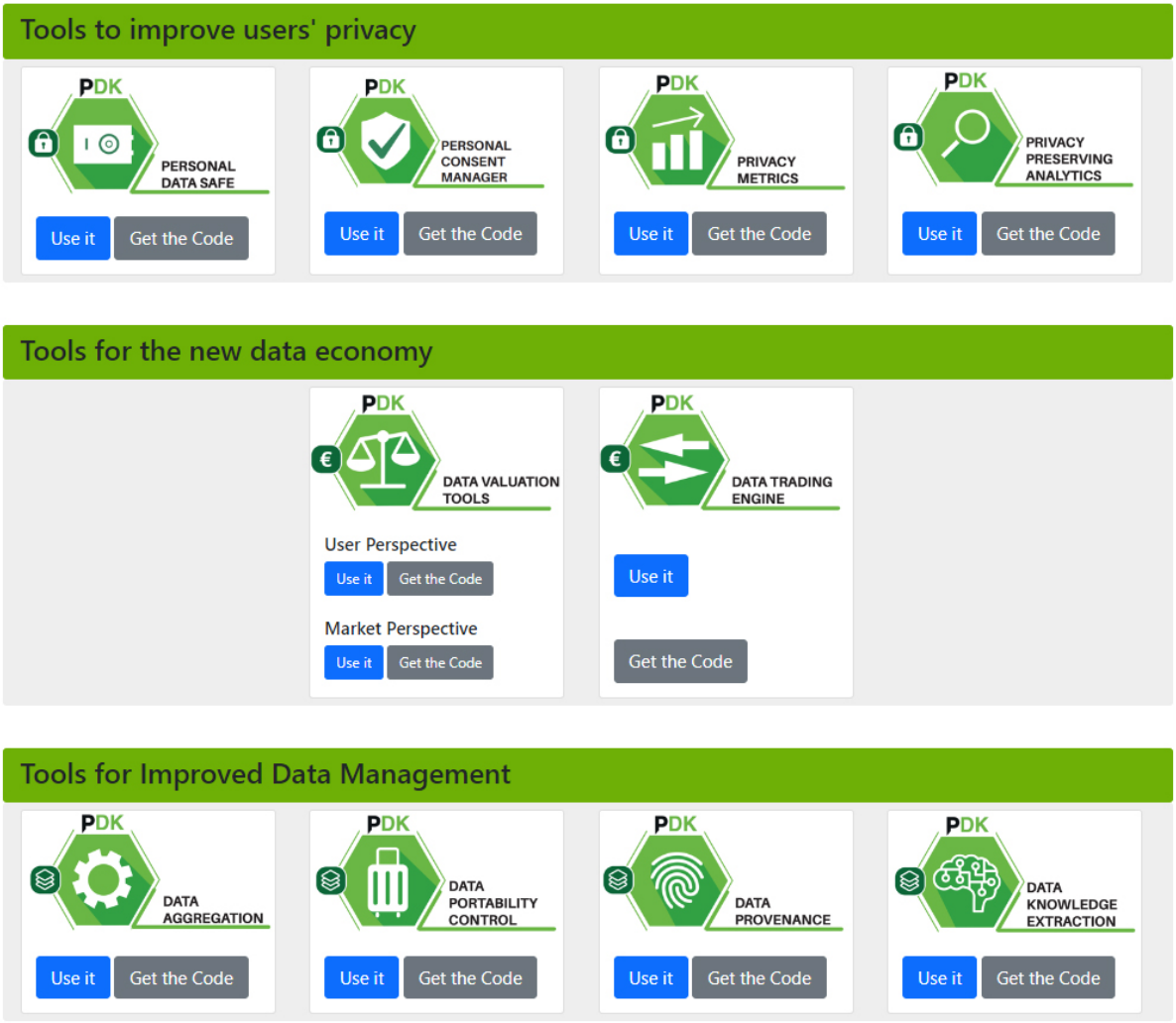

Fig 1 PIMCity online demonstration panel

Tools to improve users' privacy

These PDK modules aim to improve user privacy from various points of view. They are designed to provide users with a simple and intuitive interface and enable transparent data management. Users can use Personal Data Safe (P-DS) to securely store their personal data and eventually allow data buyers to access them through the Personal Consent Manager (P-CM). Details about data buyers can be found in the Personal Privacy Metrics (P- PM), along with information on the purpose of a data buying campaign. Finally, Personal Privacy- Preserving Analytics (P-PPA) provide data buyers access to aggregated and anonymized data by implementing anonymisation via well-known approaches such as k-anonymity, differential privacy, or z-anonymity for streams.

The **Personal Data Safe (P-DS)** is the means to store personal data in a controlled form. It implements a secure repository for the user's personal information like navigation history, contacts, preferences, personal information, etc. It gives the possibility to handle them though REST-based APIs or a web interface. Thanks to the REST APIs, the P-DS can be accessed also by other components of the PDK. The architecture of the PDK is depicted in the figure below:

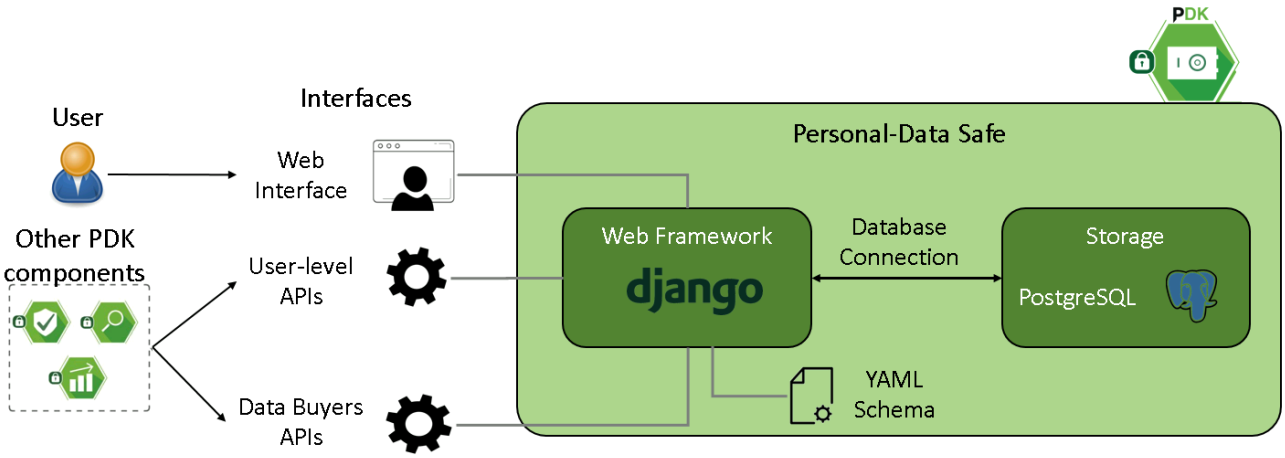

Fig 2 Personal Data Safe Architecture

P-DS reference links

How to use: https://easypims.pimcity-h2020.eu/intro-pds.html
Video: https://youtu.be/wuHDJjSZB1Y
Get Code: https://gitlab.com/pimcity/wp2/personal-data-safe

P-CM Personal Consent Manager

The primary objective of the Personal Consent Manager (P-CM) is to give the users the transparency and control over their data in a GDPR compliant way. That is, give them the possibility to decide which data can be uploaded and stored in the platform, as well as how (raw, extracted or aggregated) data can be shared with Data Buyers in exchange for value when the opportunity arises.

The P-CM is presented as a web application and a REST API, not only providing users the possibility to use the component in a user-friendly way, but also enabling developers to integrate PIMCity.

Consent Management capabilities in their products. The architecture of the PDK is depicted in the figure below. Being it a server-side module, no user interface is offered, and as such no video showcasing it is presented.

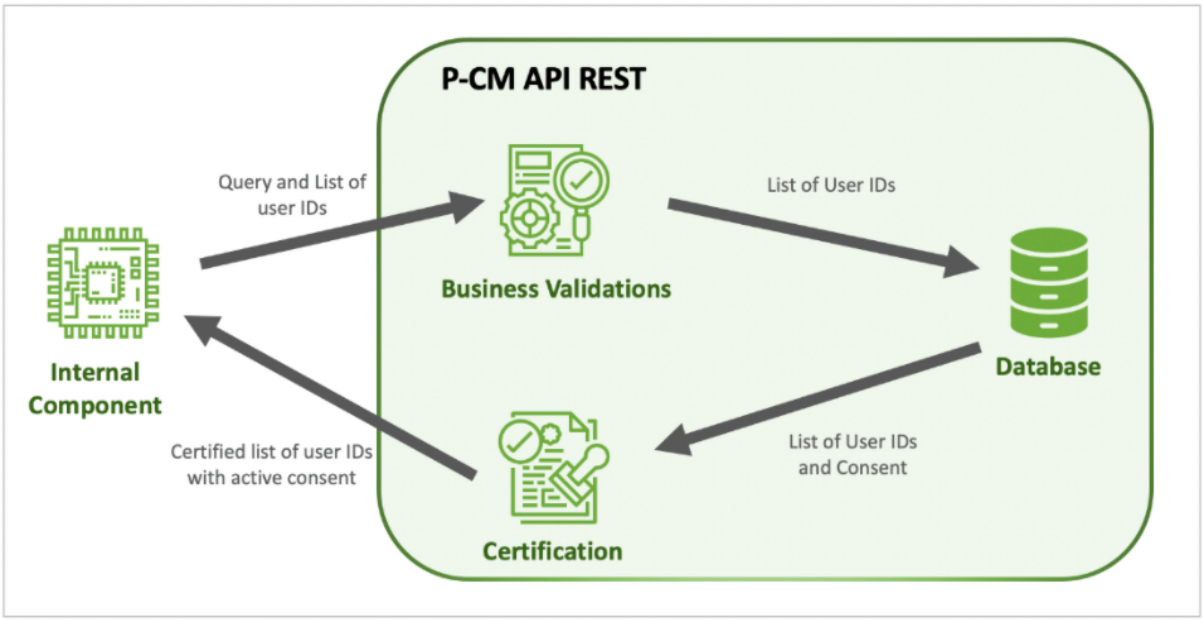

Fig 3 Personal Consent Manager architecture

P-CM reference links

Get Code: https://gitlab.com/pimcity/wp2/personal-consent-manager

P-PM Personal-Privacy Metrics

Personal Privacy Metrics (P-PM) represent the means to increase the user's awareness. This component collects, computes and shares easy-to-understand data to allow users know how a service (e.g., a data buyer) stores and manages the data, if it shares it with third parties, how secure and transparent it looks, etc.

These are all fundamental pieces of information for a user to know to take informed decisions. The PM computes this information via a standard REST interface, offering an open knowledge information system which can be queried using an open and standard platform. PMs combine information from supervised machine learning analytics, services themselves and domain experts, volunteers, and contributors. Its implementation builds on MongoDB (for the database), Python/Flask and Swagger (for the server).

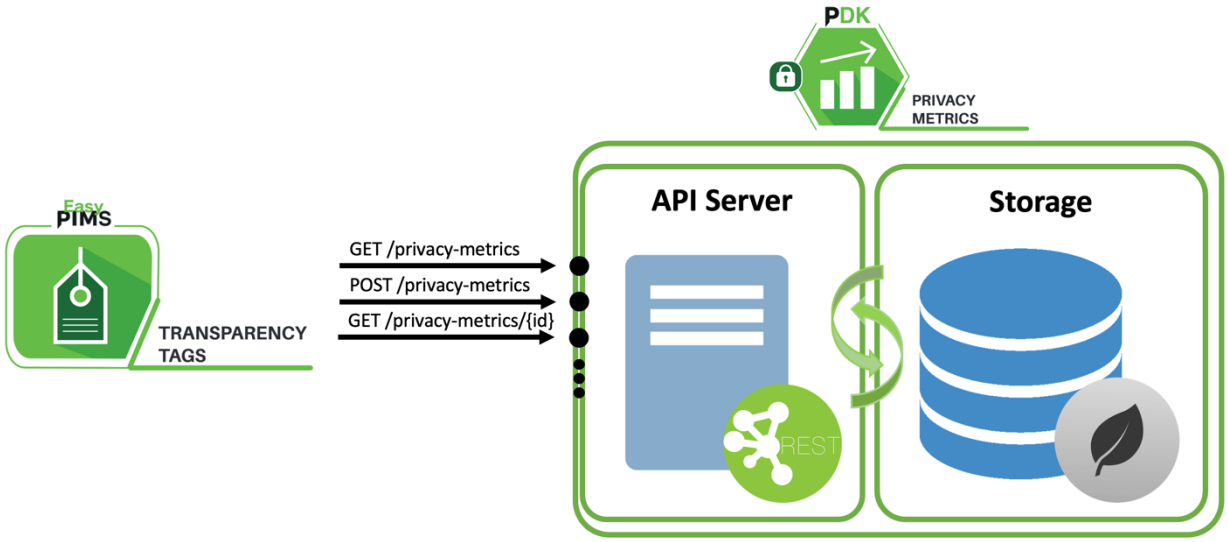

Fig 4 Personal Privacy Metrics architecture

P-PM reference links

How to use: https://easypims.pimcity-h2020.eu/intro-ppm.html
Video: https://youtu.be/SdGuTt98JRg
Get Code: https://gitlab.com/pimcity/wp2/privacy_metrics

P-PPA Personal-Privacy Preserving Analytics

The Personal Privacy Preserving Analytics (P-PPA) module has the goal of allowing data analysts and stakeholders to extract useful information from the raw data while preserving the privacy of the users whose data is in the datasets. It leverages concepts like Differential Privacy and K-Anonymity so that data can be processed and shared while guaranteeing privacy for the users

P-PPA includes a set of functionalities that allow perform data operations preserving the major privacy properties: k-anonymity, z-anonymity, differential privacy. P-PPA is capable to handle different sources of data inputs, that define which kind of privacy property is called into account: we have design solutions for tabular and batch stream, handled with PostgreSQL, MongoDB, and CSV modules, and live stream data. The figure below depicts the P-PPA architecture:

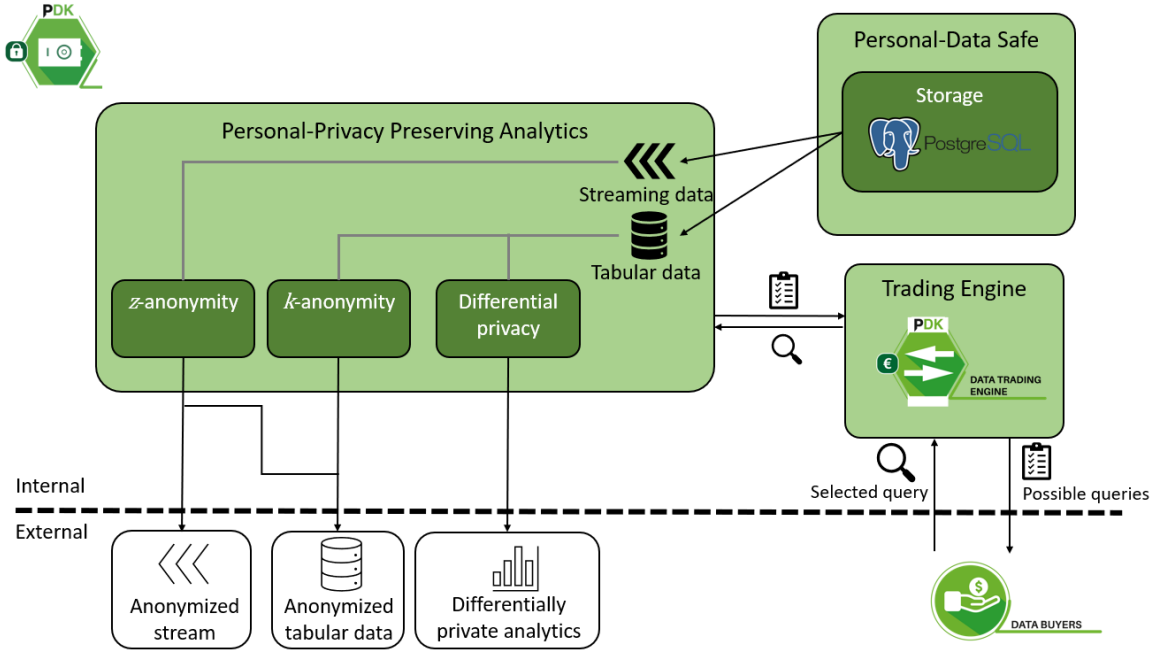

Fig 5 Privacy Preserving Analytics architecture

P-PPA reference links

How to use: https://easypims.pimcity-h2020.eu/intro-pppa.html
Video: https://youtu.be/uA7O8mFZGo4
Get Code: https://gitlab.com/pimcity/wp2/personal-privacy-preserving-analytics

Tools for the new data economy

Currently, users are not part of the data market. Conversely, they are external actors who merely provide the assets but have no influence or decision power. In this scenario, the value of end-users is determined solely by the market, i.e., the price that data buyers are willing to pay for a given end-user's data. However, in the human-centric data economy envisioned by PDK, this one-sided vision is no longer valid. Conversely, end-users must have control over their data. Hence, we come into a new scenario with two sides: the market and the users. To this end, we offer the Data Valuation Tools (DVT), which are able to derive the value of end-user data from the two perspectives mentioned above: Market and End User perspective, i.e., how much the data is worth for the buyer and for the end user, respectively. We also provide a Data Trading Engine (DTE) that can be integrated as part of the PIMS infrastructure to trade end-user data within the ecosystem.

Data Valuation Tool from the Users' perspective (DVTUP)

The Data Valuation Tool from an End-User Perspective (DVTUP) module is meant to provide estimated valuations of end-users' data for dataset sold through the marketplace as bulk data. In particular, DVTUP will provide tools for the TE to:

Provide buyers with a hint of how valuable a piece of data is for a certain type of model or even for a specific task. Calculate a fair breakdown of data transaction charges by seller, looking forward to rewarding each user proportionally to the value that each piece of data from different sellers brings to the buyer for a specific task.

In the first case, the output will be the expected accuracy the buyer would get from a specific dataset of the marketplace.

In the second case, the output will estimate the percentage of a transaction value that corresponds to each seller, and a log of data and results obtained to justify rewards paid to different sellers. For that purpose, different methodologies and algorithms will be designed and implemented to allow data marketplaces to calculate such breakdown in different ways, namely:

- Functionality 1: using traditional heuristics such as data volume or the number of sources See the paper Try Before You Buy: A practical data purchasing algorithm for real-world data marketplaces.

- Functionality 2: using a value-based data evaluation framework to test data on the specific task the buyer is intended to use it in, and breaking the payment according to a value function (e.g., accuracy yielded by a given model or algorithm) provided by the buyer. For more information, please see the paper Computing the Relative Value of Spatio-Temporal Data in Wholesale and Retail Data Marketplaces.

Functionality 1:

Let buyers test the data for their specific task, leading to a close to optimal purchasing process performance

Functionality 2:

Let the TE reward users according to the value they bring to the specific task once a transaction is closed

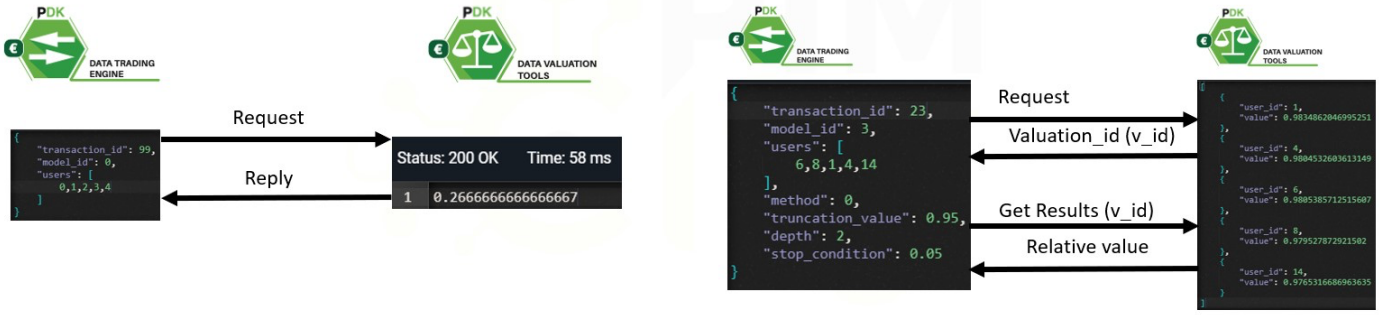

Fig 6 Data Valuation Tools users perspective schema

DVTUP reference links

How to use: https://easypims.pimcity-h2020.eu/intro-dvtup.html
Video: https://youtu.be/Wj7uQWj2pO4
Get Code: https://gitlab.com/pimcity/wp3/dvtup

The Data Valuation Tools from the market perspective (DVTMP)

The Data Valuation Tools from the market perspective (DVTMP) module developed in PIMCity is leveraging some of the most popular existing online advertising platforms to estimate the value of hundreds to thousands of audiences.

The DVTMP module aims to provide the monetary value of audiences traded on the main online advertising platforms. This serves any PIM deciding to implement the DVTMP module to have a realistic estimation of audiences' value to be traded. Since the information about values collected from these advertising platforms is based on aggregated historical pricing data, we can assert that DVTMP provides full-privacy guarantees. Moreover, a given audience's value can be obtained in real-time from the referred online advertising platform.

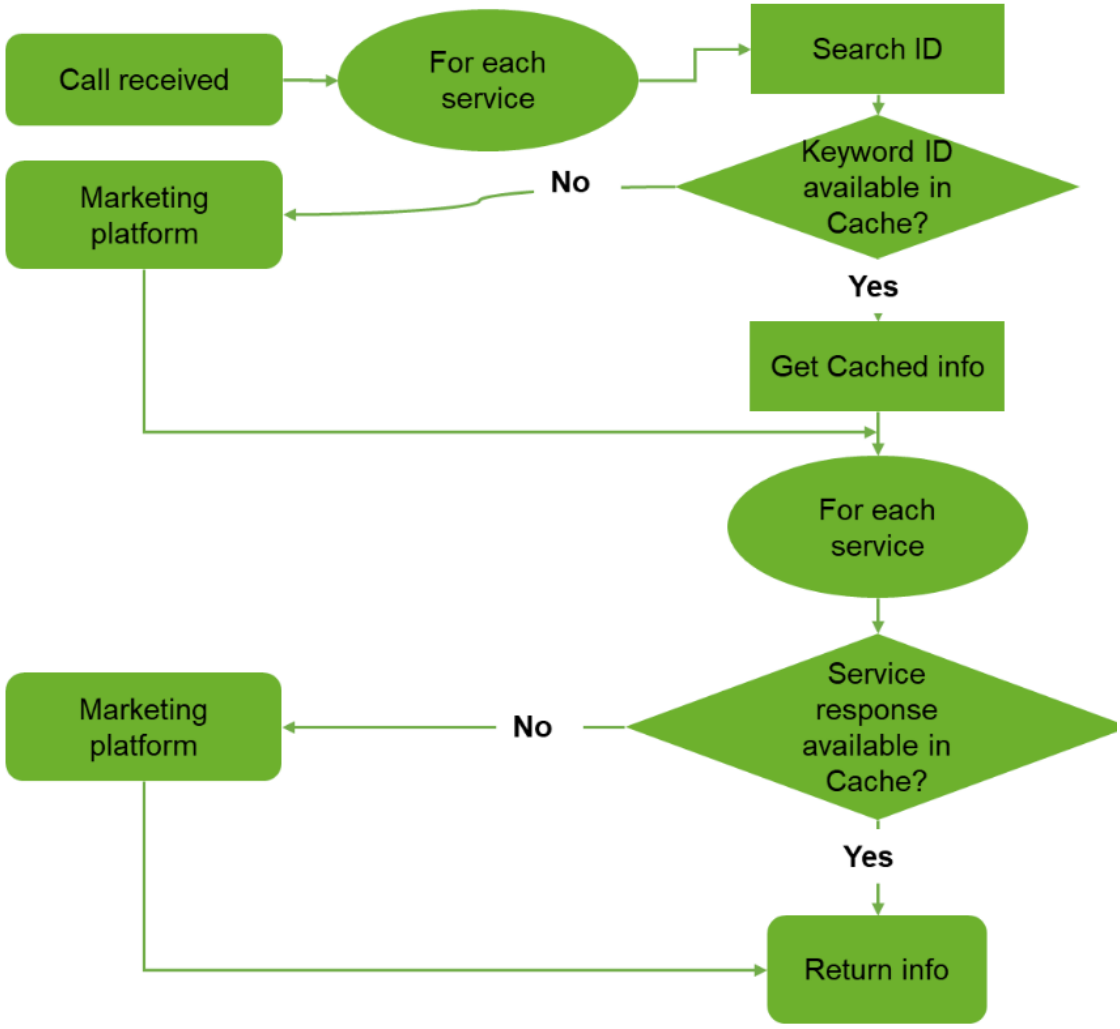

Fig 7 Data Valuation Tools from the market perspective

DTVMP reference links

How to use: https://easypims.pimcity-h2020.eu/marketplace/auValuator/demo
Video: https://www.youtube.com/watch?v=xpgpioVsALw&t=4s
Get Code: https://gitlab.com/pimcity/wp3/dvtmp

Data Trading Engine (DTE)

The Data Trading Engine (DTE) is responsible for trading the data of users registered/handled by the PIM with interested buyers. Hence, the DTE serves as a communication interface between the PIM backend and the data buyers. There is a myriad of data types that can be sold. The DTE focuses in bulk data and audience data.

Bulk data is typically bought in a non-real-time manner to receive information from a (typically large) group of users. Some examples include: a health insurance company may be interested in people's medical records; a car insurance company may be interested in the people's mobility data to understand who drives through tough roads or at higher speeds; a mortgage-issuing company may be interested in people's financial records, etc. None of these data need to be traded in real-time, and typically data buyers try to buy it in bulk.

An audience is a term used in marketing to refer to a specific group of the population defined by three parameters: location (where the user is located); demographic information (age, gender, etc.); interests (a list of interests, e.g., outdoor activities, sports, science, automobile, etc.). Most of these parameters are typically extracted from well-known taxonomies such as the one offered by IAB (which is a de-facto standard in digital marketing). Online advertising (a.k.a. digital marketing) is arguably one of the most important businesses exploiting data utilization, specifically audience data, to deliver targeted ads to users in real-time.

The DTE is exposed as a REST API which can be accessed by other components in PIMCity or external players like Data Buyers and individuals. The creation of a transaction is depicted in the figure below. Being it a server side module, no user interface is offered, and as such no video showcasing it is presented.

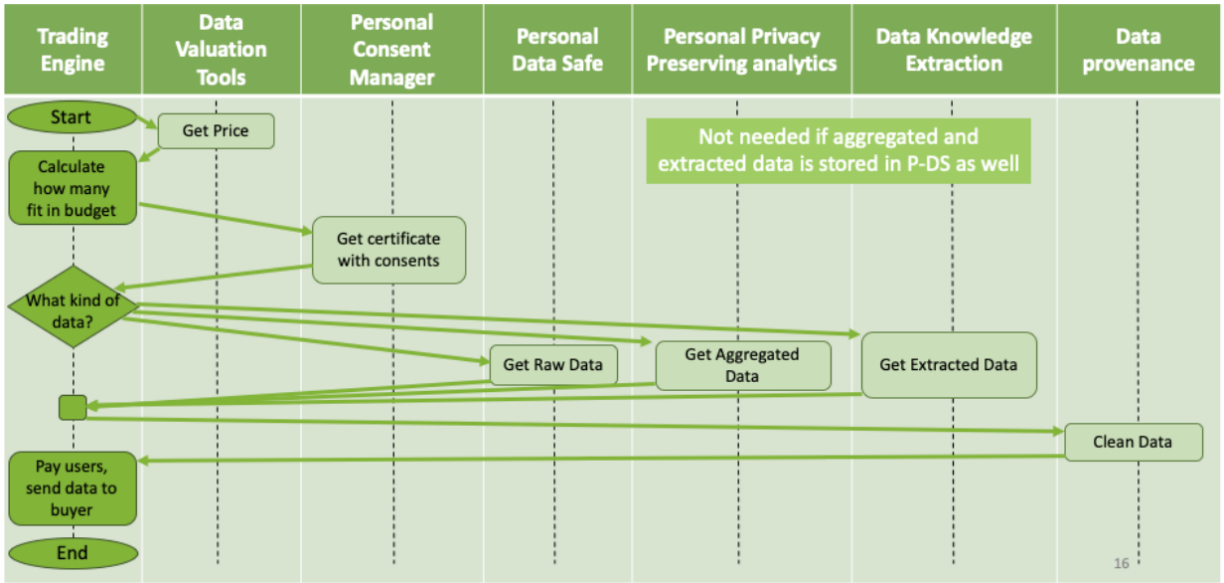

Fig 8 Data Trading Engine

DTE reference links

Get Code: https://gitlab.com/pimcity/wp3/datatradingengine

Tools for Improved Data Management

Due to the variety of devices and data sources available, it is challenging to import, process and aggregate data in a standardised, scalable, and privacy-preserving manner. To this end, we offer the Data Aggregation (DA) tool for mass insertion of personal data into a PIMS – allowing to bulk import data from large databases such those of banks or Internet Service Providers.

Portability Control (DPC) enables users to import the data directly from Facebook or Google, for example, offering filtering capabilities to define which data one is willing to import. The Data Provenance tool (DP) adds hard-to-remove watermarks to datasets to prove ownership later. Finally, the Data Knowledge Extraction (DKE) engine is an example of machine learning analytics to extract privacy-preserving models from data. It supports the creation of user profiles that contain the interests of each user as extracted from their browsing history.

Data Aggregation (DA)

The Data Aggregation (DA) tool enables data owners (for example an Internet Service Provider -ISP that hold a bulk of their users' data) to perform two important processes on their data: aggregation and anonymization. Such processes enable data owners to share these data in a privacy-preserving way.

The DA tool resides on the data owner's side and its input is the raw data that is available through the initial sources (telco data, sensor data, etc.) and it is transformed in a predefined schema / metadata model. The user (data owner) is responsible for preparing the data for processing (i.e., export from their initial source (internal database), clean them if needed, etc.). Afterwards, through the module, the user can choose the subset of the data to be aggregated / anonymized and set the related algorithmic parameters (for aggregation and anonymization).

The output is the processed (aggregated / anonymized) data that can be exported to the PIMCity marketplace through an API that the module provides. The data resides on the data owner side and the interested party can retrieve them through this API. Being it a server-side module, no user interface is offered, and as such no video showcasing it is presented.

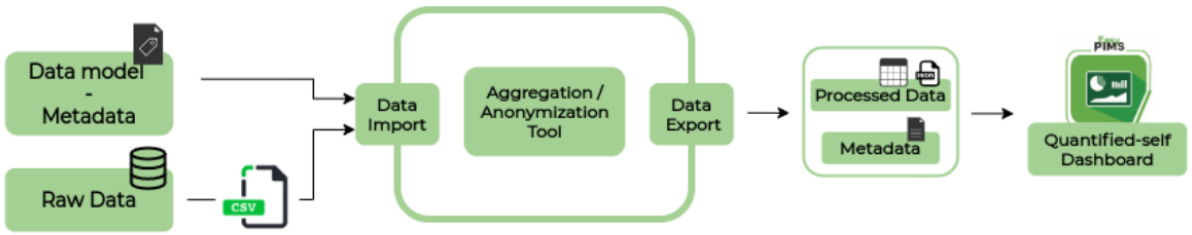

Fig 9 Data Aggregation architecture

DA reference links

Get Code: https://gitlab.com/pimcity/wp4/data-aggregation-api

Data Portability Control (DPC)

The Data Portability Control (DPC) allows users to migrate their data to new platforms, in a privacy-preserving fashion. More specifically, it incorporates the necessary tools to import data from multiple platforms (through the available Data Sources), process the data to remove sensitive information (through the Data Transformation Engine), and outport into other platforms (through the Data Export module).

The tool does not provide a dedicated UI to the users. Instead, it provides an interface in a form of a generic Control API for controlling all operations from other modules of the PIM system (e.g., the User Dashboard). The figure below depicts the DPC architecture. Being it a server-side module, no user interface is offered, and as such no video showcasing it is presented.

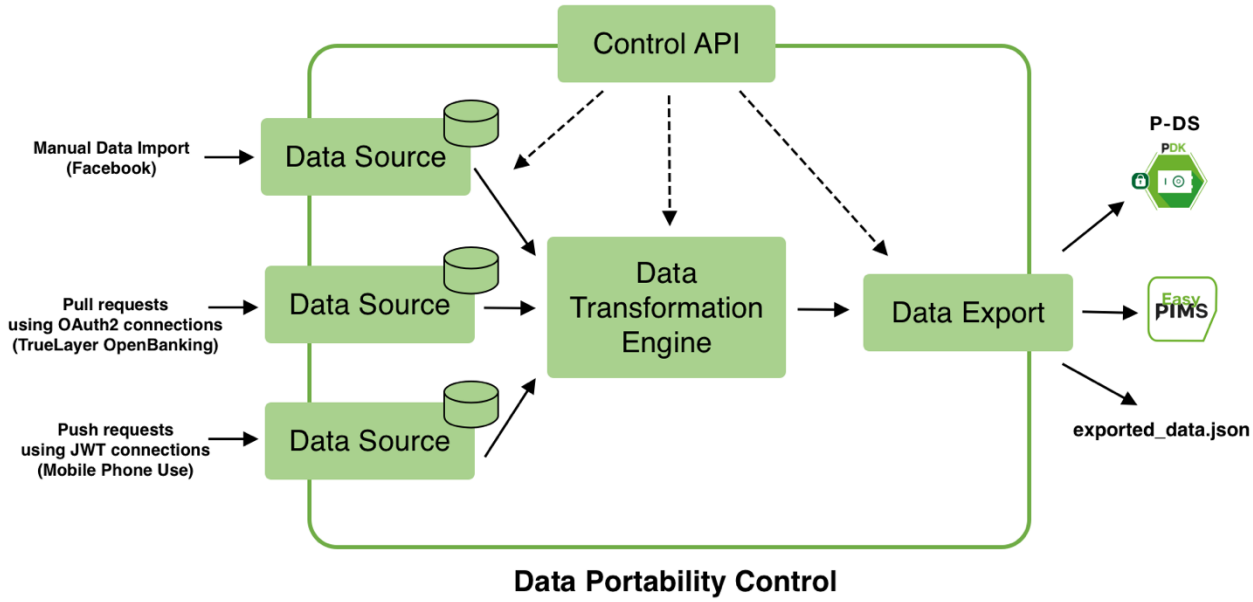

Fig 10 Data Portability Control

Reference links

Get Code: https://gitlab.com/pimcity/wp4/data-portability-control-tool

Data Provenance (DP)

The Data Provenance (DP) is a data management tool to watermark sensitive data as user web browsing history while accounting for user data ownership. It implements algorithms from the database watermarking literature (e.g., VLDB) and aims to bring new research into the area in order to use it in real world data management use cases as ours. We focus in web browsing data, namely URLs, which are a valuable piece of information about user's preferences and behavior, yet not monetizable by data owners in a decentralized manner in the real world yet (only centralized companies as Comscore exist for that).

Therefore, out tool opens a new possibility to users to sell watermarked data with the support of the Trading Engine component (out of scope in this demo and intro) so that users just need to rely on REST-based APIs or a web interface to control their data ownership. Thanks to the REST API endpoints, the DP tool can be accessed also by other components of the PDK. Internally, it uses the SpringBoot framework and will store user data on a secure PostgreSQL database as well as decentralized storage thanks to the support of IPFS (InterPlanetary File System) as middleware.

Note, in the future watermarked datasets will be encrypted with the appropriate public and or private keys, but that is out of the scope for now. The Web interface is provided by the Swagger OpenAPI tools in our deployment as a single page application.

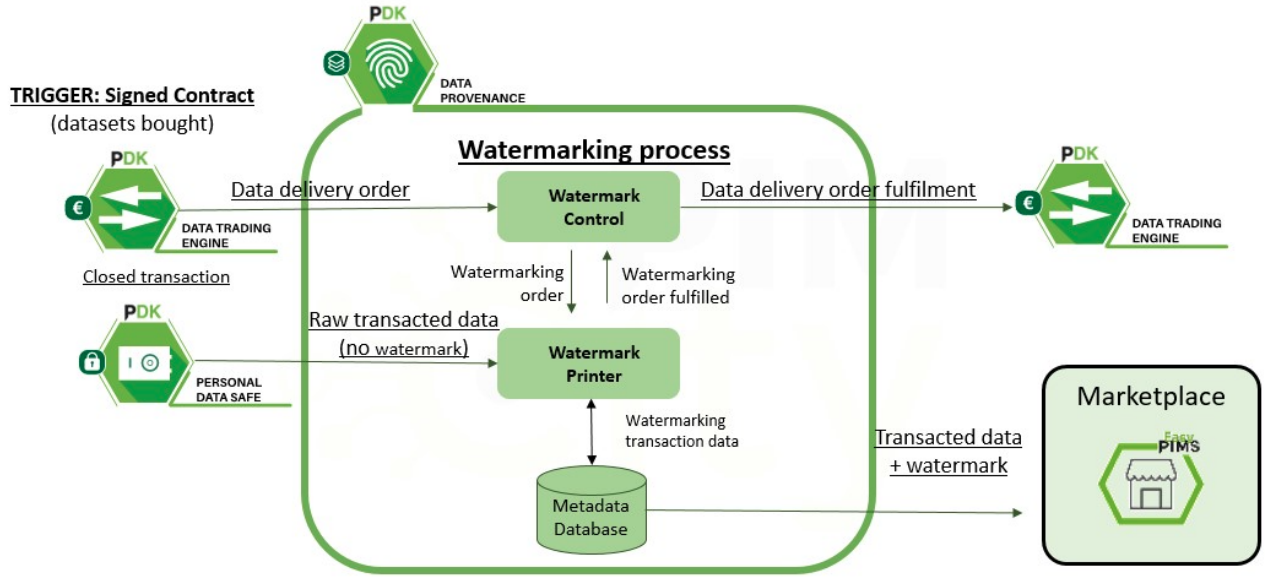

Fig 11 Data Provenance architecture

DP reference links

How to use: https://easypims.pimcity-h2020.eu/intro-provenance.html
Video: https://youtu.be/iRuDRYrPsZw
Get Code: https://gitlab.com/pimcity/wp4/data-provenance

DKE: User Profiling System

The User Profiling System (Part of the DKE) is able to automatically generate user profiles from the sequences of hosts visited by users. The User Profiling System is developed in Python and makes use of the Gensim library to train a skipgram model for the network hosts

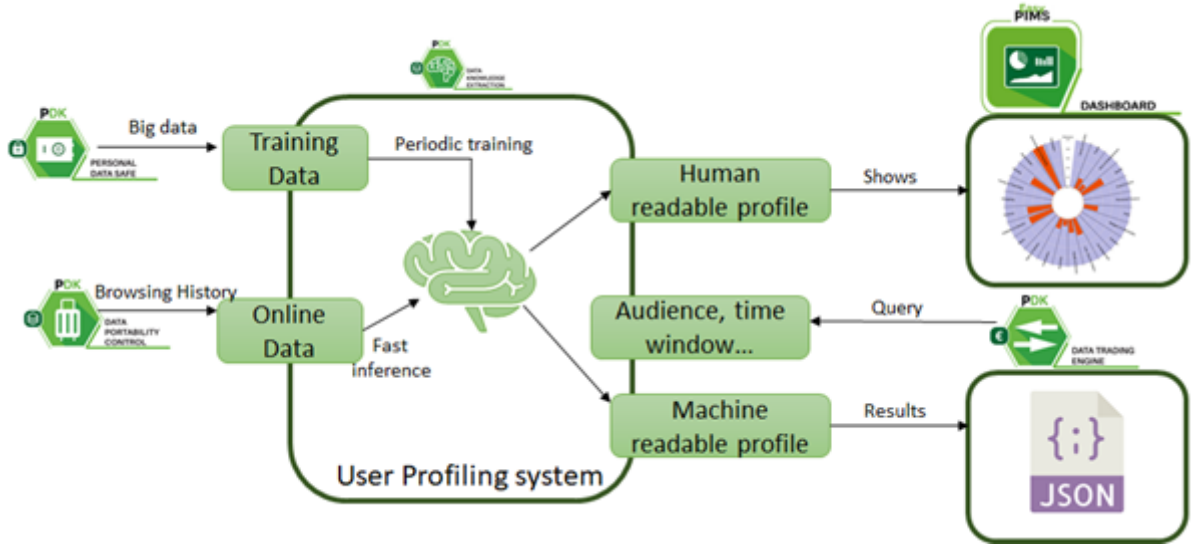

Fig 12 Data Knowledge Extraction architecture

DKE reference links

How to use: https://easypims.pimcity-h2020.eu/intro-dke.html
Video: https://youtu.be/UxLNeVIGcY8
Get Code: https://gitlab.com/pimcity/wp4/userprofilingsystem

PIMCity project generic training materials

The materials developed to make a general presentation of the project are all available on-line and are as follows:

- Video presentation of the project: https://youtu.be/Mb9r37T5PZw

- PIMCity Project Presentation (PPT): https://pimcity.eu/docs/PIMCity_Presentation.pptx

- PIMCity presentation brochure (PDF): https://pimcity.eu/docs/PIMCity_Brochure.pdf

- EasyPIMS presentation Brochure (PDF): https://pimcity.eu/docs/EasyPIMS_Brochure.pdf

- General project presentation article (PDF): https://pimcity.eu/docs/IEEE_PIMS_PDK.pdf

Conclusions

This document contains a summary of the main basic technology components developed in the PIMCITY project as components of the PDK. These components are exploitable outcomes on their own.
This documentation has been used to implement the different demonstrators developed in the last stage of the project.

The materials collected here have already been successfully used in some of the seminars and presentations addressed to companies and developers to inform the attendees about the concepts and programs that support the developed tools.