EDITORS: Munindar P. Singh, singh@ncsu.edu
Pradeep Murukannaiah, p.k.murukannaiah@tudelft.nl

DEPARTMENT: INTERNET ETHICS

A PIMS Development Kit for New Personal Data Platforms

Nikhil Jha, Martino Trevisan, Luca Vassio, and Marco Mellia, *Politecnico di Torino, 10129, Turin, Italy*

Stefano Traverso, *Ermes CyberSecurity, 10146, Turin, Italy*

Alvaro Garcia-Recuero and Nikolaos Laoutaris, *IMDEA Networks Institute, 28918, Leganés, Spain*

Amir Mehrjoo and Santiago Andrés Azcoitia, *IMDEA Networks Institute, 28918, Leganés, Spain and Universidad Carlos III de Madrid, 28903, Getafe, Spain*

Ruben Cuevas Rumin, *Universidad Carlos III de Madrid, 28903, Getafe, Spain*

Kleomenis Katevas, Panagiotis Papadopoulos, and Nicolas Kourtellis, *Telefonica Research, 08019, Barcelona, Spain*

Roberto Gonzalez, *NEC Labs Europe, 69115, Heidelberg, Germany*

Xavi Olivares and George-Marios Kalatzantonakis-Jullien, *LSTECH ESPANA SL, 08005, Barcelona, Spain*

The web ecosystem is based on a market where stakeholders collect and sell personal data, but nowadays users expect stronger guarantees of transparency and privacy. With the PIMCity personal information management system (PIMS) development kit, we provide an open-source development kit for building PIMSs to foster the development of open and user-centric data markets.

In today's data-driven economy, the amount of data a company holds has a direct and nontrivial impact on its overall market valuation. Data are catalyzing not only business, but also governance and everyday life, across all sectors, regions, time scales, economic, and political systems worldwide. Online advertising and marketing have driven developments in this space, transforming a decades-old industry and creating some of the biggest businesses (and in a few cases, controversies) of our time. In fact, the online advertising industry is breaking records year after year, both in terms of growth and overall value.

However, online advertising is just the tip of the iceberg. Data are being sought and offered in a wide range of applications, and data-driven decision making is having a significant impact across a variety of sectors. According to a large-scale 2016 study by McKinsey,[1] the numbers for the potential of data-driven decision making are staggering, even by the most conservative estimates.

In some aspects, this economy is primitive, the source of value—or raw material—are the users, and they have no choice but to give away their goods (data) to a very few companies against whom they have no bargaining power. In exchange for their goods, users receive a range of services, some of which are now essential to everyone's digital life: web search, connecting with other people, shopping, etc. As a result, users cannot really opt out and can only continue to give away their data without being able to negotiate compensation. This is not a market. It is more like the colonial economy, where peasants had no choice but to work for the colonists, without any bargaining power whatsoever.

This situation has sparked intense debates on various issues, including privacy, discrimination and bias, manipulation of public opinion, and spread of fake news,[2] competition and monopolisation, automation and its impact on unemployment and economic inequality.[3]

1089-7801 © 2022 IEEE
Digital Object Identifier 10.1109/MIC.2022.3157356
Date of current version 21 June 2022.

Personal information management systems (PIMSs), also called personal data banks or personal data vaults, are a promising alternative to the uncontrolled collection, processing, and use of people's data. A PIMS can be thought of as a software interoperability layer between end-users and data services, responsible for ensuring that data are passed from the former to the latter in a controlled manner. However, PIMSs so far are struggling to succeed due to both the complexity of creating a fully fledged solution, and the bootstrapping problem.

In this article, after reviewing the available PIMS, we present and discuss the new PIMS development Kit—or PIMS development kit (PDK)—whose main goal is to speed up the creation, testing, and deployment of PIMSs, limiting the burden of creating a solution from scratch.

CURRENT SOLUTIONS FOR PERSONAL DATA MANAGEMENT

The first countermeasures against the collection of user data were solutions to block online advertisements and trackers, usually implemented via browser plugins. AdBlock Plus and Ghostery are notable examples. In response, services have attempted to circumvent blocking with a variety of more sophisticated tracking techniques, fuelling an arms race.

Recently, several technological solutions and business models have emerged to balance the abovementioned tensions: PIMSs. They look to empower individuals to take control of their personal data. For that purpose, they are building capability to let users collect their personal information from other sources (e.g., banks or internet service providers); exercise their erasure and modification rights; manage cookie, privacy, and access permissions settings; manage consent for sharing personal data; monetize data by allowing users to receive the corresponding payments for their sharing.

We identify 18 existing PIMSs platforms that offer the ability to trade personal data. Most PIMSs focus on collecting and managing personal data for marketing-related purposes. Some specialize in using their data for targeted marketing surveys and rewarding users for filling in questionnaires. As supplemental material, we report the complete list of the 18 surveyed PIMSs.

In terms of architecture, PIMSs are usually decentralized platforms that use users' devices to store information. They may rely on blockchain solutions to provide an additional layer of security and meet demanding regulatory requirements. In terms of data trading, nine of the 18 surveyed PIMSs focus on consent management and data sharing, and therefore, do not offer specific marketplace or data pricing features. Those that offer such features typically help users set a fair price for their data: they manage buyers' bids, advise sellers on actual prices, or adjust prices based on buyers' purposes. Finally, more than half of the PIMSs leverages their own cryptocurrency to process payments.

Most of PIMSs provide as open-source some of their components—for example, Digi.me, Airbloc, Meeco, have a public GitHub repository. However, each of the current PIMSs targets a specific use case and envisions a precise business model. This limits their applicability and acceptance, as no single platform can cover the diverse scenarios of the current data economy. Another obstacle to PIMS scaling is the need for earning the *trust* of users for them to let the PIMS manage their personal data on the Internet, which is not trivial in the current data-for-services economy.

PIMCITY PDK: CHALLENGES AND DESIGN PRINCIPLES

To unlock the potential of data-driven decision making, as part of the EU-funded PIMCity project, we have designed, developed, and validated a set of reusable, flexible, open, and user-friendly components in the form of a PDK. Being aware of the complex and nonstandard definition of PIMS, our goal is to provide a modular approach that can be flexibly improved and refined as needed. The PDK offers the ability to rapidly develop new PIMS solutions and easily experiment with possible alternatives. We have carefully developed a bottom-up methodology that involves all stakeholders (including advertisers and end-users) at all stages, from design to development to large-scale demonstration and going to market. We strongly believe that an open market for data will only thrive if we stop the arms race between users and services.

As a first tangible result, we offer the PDK to commodities the creation complexity of PIMSs. This PDK lowers the barriers for companies to enter the web data market. The main challenges in designing and developing the PDK can be summarised as follows.

User-Centric model

Implementing a user-centric data ecosystem is the biggest challenge of the PDK. A user-centric data economy requires that individuals are compensated for their data in proportion to the overall economic benefits. Then, what is a reasonable price for data? Even though PIMS users and data sellers are usually in charge of setting this price, they do not know what a fair price should be. To this end, the PDK offers a data valuation framework

backed by state-of-the-art research in the field.[4] On the one hand, these data valuation tools (D-VT) allow one to estimate the value of their data once offered on a marketplace, i.e., how much my data is worth. On the other hand, they redistribute the revenues among users whose data was traded on the market, i.e., what is the fair share of revenues my data shall be compensated.

Interoperability

PIMCity architecture allows users to integrate new data sources and connect them to new services. This is a fundamental property to build trust in any PIMS. For this, we offer predefined modules to import data from common sources (e.g., Facebook and Google exported data, the open banking protocol) into a common personal data safe (P-DS) module, the "personal data bank," which allows a user to store and manage their data, and offer them to an open marketplace.

Interoperability is the biggest advantage offered by the PDK and at the same time, a great challenge, because it requires a process of standardization of consent mechanisms, formats, and semantics. All PDK components provide REST-APIs, which we document using the Open APIs specifications to enable seamless integration. This enables communication and interaction between them and facilitates integration with existing PIMSs as well as the design and development of new ones.

Open-Source Software

We deem open-source software a means to achieve transparency and user trust. Although maintaining a (large) open-source project is challenging in terms of code maintenance and long-term support, it allows us to collect feedback, bugs, feature requests, and ultimately measure the success of the PDK. The PDK is open-source and available online on the GitLab Project of PIMCity.[5] We encourage its use and invite the community to test and support the project. We use the GitLab collaboration features as a forum for tracking issues, discussing bugs, requesting new features, and providing user support.

PDK IN DETAILS

In the PDK, we design and develop generic components that offer fundamental functionalities for PIMS. We release them as SDKs to streamline PIMS development and integration. We identify three coarse areas, in which we group those elemental blocks that offer basic functionalities and sketch them in Figure 1.

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 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 anonymization via well-known approaches, such as k-anonymity,[6] differential privacy, [7] or z-anonymity for streams.[8]

Tools for a User-Centric 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 the following two sides: the market and the users. To this end, we offer the D-VT, 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 (D-TE) that can be integrated as part of the PIMS infrastructure to trade end-user data within the ecosystem.

Tools for 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 us to bulk-import data from large databases such those of banks or Internet service providers. Data 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

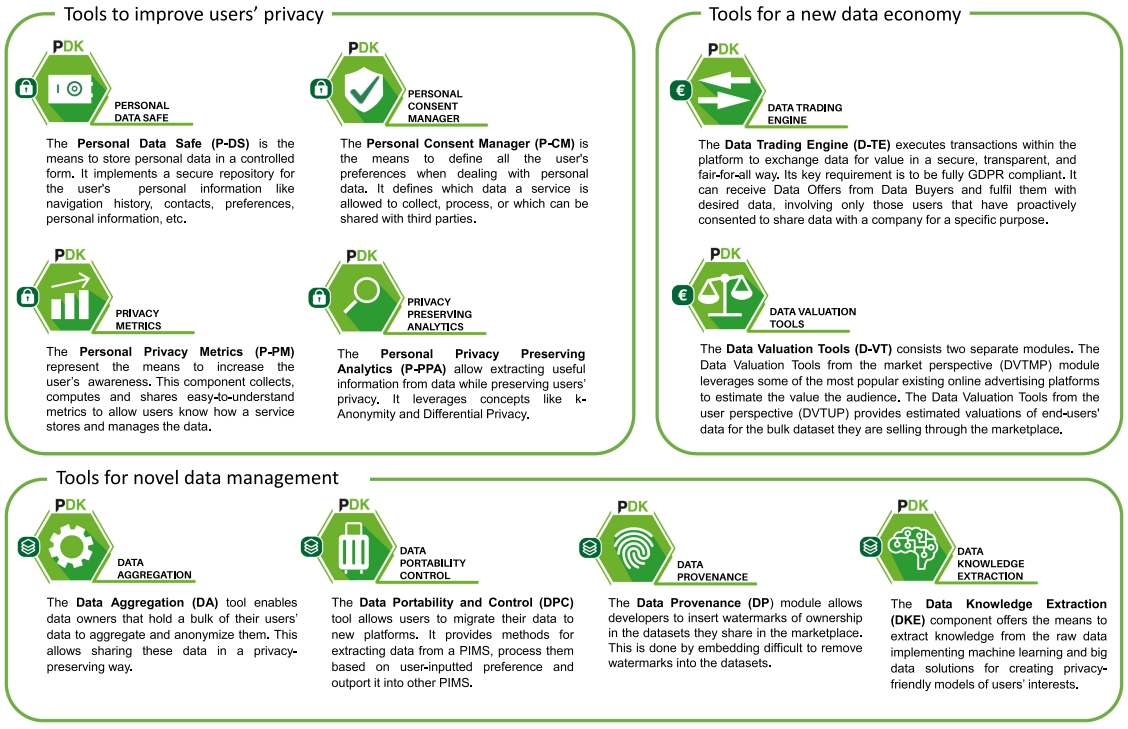


FIGURE 1. PDK components for privacy-enhanced, data management, and tools for data economy.

creation of user profiles that contain the interests of each user as extracted from their browsing history.

USE CASES AND APPLICATIONS

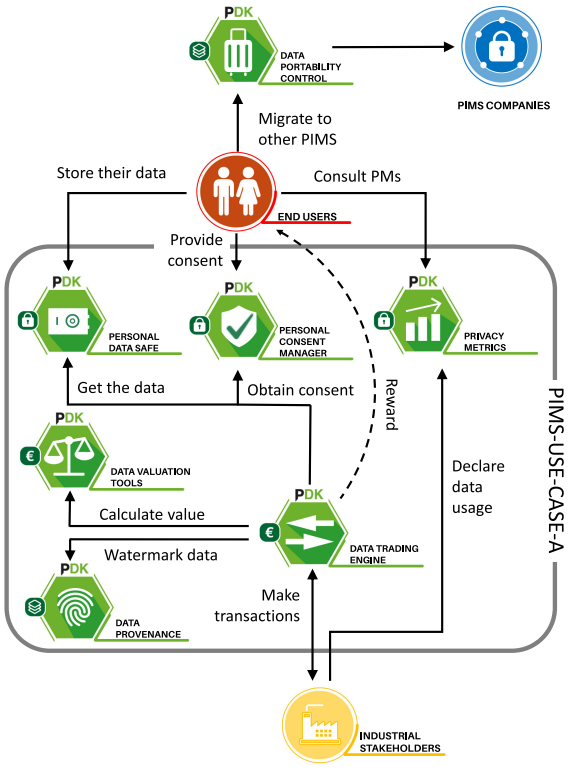
Here, we discuss two possibilities that we consider common use cases for the PDK.

Fully Fledged PIMS

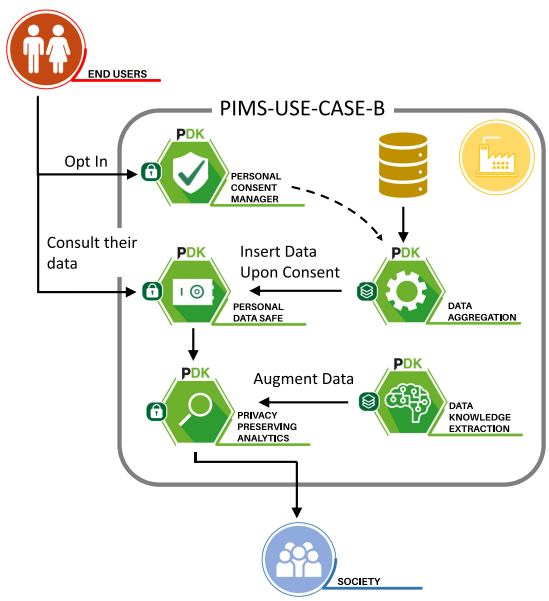
Combining our PDK modules makes it possible to build a new PIMS prototype without developing all the applications from scratch. In Figure 2(a), we show how the modules work together. Each user can store their data in the P-DS. This allows them to have structured, well-organized information about the data they provide to the system. With the help of the DPC, they can import/export their data from/to another PIMS company. Through the P-CM, the user can specify what types of data they are willing to share, to what class of data buyers, and in what form (raw, aggregated). The DP module can watermark the datasets before they are sold through the D-TE to keep the ownership of the data verifiable in a healthy data economy model. When a data buyer is interested in the users' data, the D-TE handles the request and operations, calculates the data value with the D-VT, collects users' consent on the P-CM, and offers the user a fair compensation. With this in place, any user can consult the easy-to-understand P-PM to learn about the purpose of the data purchase. This makes the user the main actor with complete control of their data and its use in the open marketplace.

IMS for Societal Benefit

Illustrated in Figure 2(b), we consider a PIMS in the premises of a company that holds personal data as a consequence of its business. For instance, a telecommunication provider with access to customer location data, or an online store with customer purchase history. The use case encourages users to share their personal data in exchange for a reward from the company, which can be statically determined (e.g., a discount on the monthly subscription) or dynamically defined using the DVT (not shown in the figure). Customers can opt-in using the P-CM, giving the company the right to share their data with third parties. Upon consent, the P-DS stores users' data using the DA module to perform a bulk transfer from the company's systems. The P-PPA allows third parties to perform privacy-preserving queries that aggregate data from multiple customers and obtain an anonymized version of a portion of the dataset, protecting the identity of individual customers. Finally, the DKE can enrich the raw data by creating user profiles. Interested stakeholders can access the system to collect anonymized data and perform their own analytics.



(a) Use Case A: Users store data in a PIMS and get a reward when a Data Buyer purchases data.



(b) Use Case B: A company's users can opt in to share the data with third parties, e.g., for societal benefit.

FIGURE 2. Use cases for the PDK. (a) Use Case A: Users store data in a PIMS and get a reward when a data buyer purchases data. (b) Use Case B: A company's users can opt in to share the data with third parties, e.g., for societal benefit.

DISCUSSION AND ONGOING DEPLOYMENT INITIATIVES

With the PDK, we simplify experimentation enabling the prototyping of new user-centric marketplaces and fostering a new data economy where users are at the centre and have complete control over their data. The PDK includes tools for managing consent and personal data, and for creating marketplaces.

The development of the PDK was a two-year effort for the PIMCity team. We designed it to identify PIMS basic functionalities and offer independent components that are easy to integrate and use. A PIMS must include many and diverse functionalities, and therefore, it is not trivial to find a satisfactory level of component integration and interoperability. The biggest challenge (and the most important lesson we learned) is that the current ubiquity of data in our lives makes creating generic components a complex task. Data come in a variety of formats (location, health, and browsing data are just a few examples) and from heterogeneous players, which are typically not cooperative and do not offer standard means to export data from their platforms but instead offer cumbersome and time-consuming means to obtain your personal data. Therefore, PIMSs must be flexible and constantly updated.

To disseminate our work to users and enterprises, we are currently developing two pilot projects that demonstrate how the PDK simplifies the development of complete solutions. The first pilot is the EasyPIMS platform, a PIMS for end-users where anyone can offer their data and receive rewards from companies interested in running data collection campaigns. The second pilot project is devoted to testing the PDK in a business-to-business scenario, involving companies interested in products that combine security and privacy protection with user education and awareness.

ACKNOWLEDGMENTS

This work was supported in part by the European Union's Horizon 2020 research and innovation programme under Grant 871370 (PIMCity) and in part by the SmartData@PoliTO center for big data technologies.

REFERENCES

1. M. Analytics, "The age of analytics: Competing in a data-driven world," McKinsey & Company, San Francisco, CA, USA, 2016. [Online]. Available: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

2. C. Ireton and J. Posetti, *Journalism, Fake News & Disinformation: Handbook for Journalism Education and Training*. Paris, France: Unesco Publishing, 2018.
3. J. Lanier, *Who Owns The Future?*. New York, NY, USA: Simon and Schuster, 2014.
4. J. Pei, "Data pricing—From economics to data science," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 3553–3554.
5. "PIMCity PDK online." Accessed: Oct. 01, 2021. [Online]. Available: https://gitlab.com/pimcity/
6. P. Samarati, "Protecting respondents identities in microdata release," *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 6, pp. 1010–1027, 2001.
7. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3/4, pp. 211–407, 2014.
8. N. Jha, T. Favale, L. Vassio, M. Trevisan, and M. Mellia, "Z-anonymity: Zero-delay anonymization for data streams," in *Proc. IEEE Int. Conf. Big Data*, 2020, pp. 3996–4005.

NIKHIL JHA is with the Politecnico di Torino, 10129, Turin, Italy. Jha received his M.Sc. degree in telecommunication engineering from Politecnico di Torino in 2019. Contact him at nikhil.jha@polito.it.

MARTINO TREVISAN is with the Politecnico di Torino, 10129, Turin, Italy. Contact him at martino.trevisan@polito.it.

LUCA VASSIO is with the Politecnico di Torino, 10129, Turin, Italy. Contact him at luca.vassio@polito.it.

MARCO MELLIA is with the Politecnico di Torino, 10129, Turin, Italy. He is a Fellow of the IEEE. Contact him at marco.mellia@polito.it.

STEFANO TRAVERSO is with Ermes CyberSecurity, 10146, Turin, Italy. Contact him at s.traverso@ermes.company.

ALVARO GARCIA-RECUERO is with IMDEA Networks Institute, 28918, Leganés, Spain. Contact him at alvaro.garcia@imdea.org.

NIKOLAOS LAOUTARIS is with IMDEA Networks Institute, 28918, Leganés, Spain. Contact him at nikolaos.laoutaris@imdea.org.

AMIR MEHRJOO is with IMDEA Networks Institute, 28918, Leganés, Spain, and Universidad Carlos III de Madrid, 28903, Getafe, Spain. Contact him at amir.mehrjoo@imdea.org.

SANTIAGO ANDRÉS AZCOITIA is with IMDEA Networks Institute, 28918, Leganés, Spain, and Universidad Carlos III de Madrid, 28903, Getafe, Spain. Contact him at santiago.azcoitia@imdea.org.

RUBEN CUEVAS RUMIN is with Universidad Carlos III de Madrid, 28903, Getafe, Spain. Contact him at rcuevas@it.uc3m.es.

KLEOMENIS KATEVAS is with Telefonica Research, 08019, Barcelona, Spain. Contact him at kleomenis.katevas@telefonica.com.

PANAGIOTIS PAPADOPOULOS is with Telefonica Research, 08019, Barcelona, Spain. Contact him at panagiotis.papadop@telefonica.com.

NICOLAS KOURTELLIS is with Telefonica Research, 08019, Barcelona, Spain. Contact him at nicolas.kourtellis@telefonica.com.

ROBERTO GONZALEZ is with NEC Labs Europe, 69115, Heidelberg, Germany. Contact him at roberto.gonzalez@neclab.eu.

XAVI OLIVARES is with LSTECH ESPANA SL, 08005, Barcelona, Spain. Contact him at xavi.oliza@lstech.io.

GEORGE-MARIOS KALATZANTONAKIS-JULLIEN is with LSTECH ESPANA SL, 08005, Barcelona, Spain. Contact him at george@lstech.io.