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

DELIVERABLE D8.5 **OPEN RESEARCH DATA PILOT**

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)	

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

Approvals

	Name	Entity	Date	Visa
WP Leader	Marco Mellia	POLITO	12/08/2022	
Author	Tiziana Rolando	POLITO	12/08/22	
Author	Marco Mellia	POLITO	17/08/22	
Reviewer	Enzo Marquet	KUL	29/08/22	
Final editor	Marco Mellia	POLITO	31/08/02	☑

Document history

Revision	Date	Modification
Version 1	12/08/2022	Initial Version
Version 2	17/08/2022	Update content and inserted additional material
Version 3	31/08/2022	Reviewed by KUL and finalized as final version

List of abbreviations and acronyms

Abbreviation	Meaning
G.A.	Grant Agreement
CA	Consortium Agreement
GA	General Assembly
PB	Project Board
PC	Project Coordinator
PrO	Project Office
IR	Interim Reports

Disclaimer

The information, documentation and figures available in this deliverable are written by the PIMCity Consortium partners under EC co-financing and does not necessarily reflect the view of the European Commission. The information in this document is provided "as is", and no guarantee or warranty is given that the information fits any particular purpose. The user uses the information at its sole risk and liability.

Executive Summary

The PIMCity project participates in the Open Research Data Pilot of the European Commission. Projects participating in this Pilot are required to:

- Develop (and keep up-to-date) a Data Management Plan (DMP).

- Deposit their data in a research data repository.

- Ensure third parties can freely access, mine, exploit, reproduce and disseminate their data.

- Provide related information and identify (or provide) the tools needed to use the raw data to validate their research.

This report details the steps taken by the Consortium towards achieving the aforementioned goals.

Table of Contents

1. Abstract

This report details the management of datasets created by the PIMCity project. In particular, we provide a summary of data created and used by the project and then we describe the measures taken towards making the datasets findable, accessible, interoperable, and reusable, i.e., according to the FAIR principles. We also discuss the ethical aspects of the datasets used and created, which is particularly critical given the type of personal data collected during the project.

2. Data summary

The PIMCity project collects and/or generates at least four broad categories of data as provided in the table below. The information provided further in the DMP (Data Management Plan – detailed in D7.6) clarifies this information further. Particularly, the DMP elaborates on the management of these broad categories of data and reveals in detail the relation of each dataset with the PIMCity project's exploitable outputs which allows gaining a clear understanding of the limitations of the OA (Open Access) as well as the impact on the exploitation and dissemination strategies. The table below summarizes the main type of artifact produced by the project, specifying the format of the information, the access type, and the details on who took care of organizing the material.

Category	Format	Means for OA	Curation and cost allocation
Documents and dissemination materials: Includes deliverables, reports, demonstrations, dissemination and communication	Common text, image or video formats (.pdf, .docx, .jpeg, .mov, .avi, etc.)	Self-archive on website; green scientific publications + OPENAIRE repositories	Technical coordinator (T8.2); dissemination manager (T6.1); peer review: scientific journal panels
Computer software: including software applications (in binary form), libraries in the form of SDKs, plugins and respective source code	Binary format, ZIP archives; Source code in common files such as C, CPP, etc.	Open repositories on GitLab	Technical coordinator (T8.2); innovation manager: exploitability and license schemes (T6.4)
Research data and metadata: materials and datasets resulting from the implementation of the developments; metadata and configuration files; bug logs and feedback logs; developer internal documentation; evaluation and opinions	Log files; text files using (.pdf, .docx, .xls, etc.)	green scientific publications + OPENAIRE repositories	Innovation manager: exploitability (T6.4); data manager: anonymization of evaluation questionnaires and opinions; conditions of pre-existent data (T7.1); dissemination manager (T6.1)
Data for evaluation: consists in materials or datasets generated or collected by the project used for evaluation purposes	Log files; files using (csv, .docx, .xls, etc.)	Green scientific publications + OPENAIRE repositories on Zenodo	Data manager: conditions of pre-existent data (T7.1); dissemination manager (T6.1)

Overall, PIMCity project partners generated a significant amount of research data with potential for re-use and verification. As identified in the PIMCity project proposal, in compliance with Responsible Research and Innovation on Open Access (further as OA), default policy is to make its data publicly available through public copyright licenses (e.g. Creative Commons, etc.). A possible solution can be, among others, archiving it on the project website and OPENAIRE compliant repositories. However, with regard to certain data, PIMCity partners may need to apply OA restrictions. The latter may stem from, including but not limited, (i) confidentiality and intellectual property protection of certain deliverables, datasets and outputs (e.g. underlying algorithmic and methods susceptible of being patented); (ii) protection of personal data of persons involved in feedback collections; (iii) protection of imported rights of pre-existent, non-public datasets.

In particular, for PIMCity the protection of personal data is of utmost importance. Given the project focuses on the collection and monetization of personal data from end users, we cannot, by any means, make this data available as OA unless explicitly permitted when collecting such data.

3. FAIR Data

The Consortium is compliant to the OPEN RESEARCH DATA PILOT as:

All papers are Open access and made available through the project website at https://www.pimcity-h2020.eu/dissemination/publications/ and through the continuous reporting tool offered by the EC at https://ec.europa.eu/research/participants/grants-app/reporting/DLV-871370.

The list of papers is available also in D6.6 - PIMCity outreach and engagement strategy and final dissemination report.

All datasets are accessible according to the FAIR principles. For papers that have open datasets (even if a limited number of dataset are available due to the type of research that involves personal data), each partner Uploaded/linked them to the ZENODO PIMCity page https://zenodo.org/communities/pimcity via the link https://zenodo.org/deposit/new?c=pimcity for upload.

4. Open access

PIMCity partners recognise that making research results more accessible contributes to better and more efficient science and to innovation in both public and private sectors. In relation to this, the default policy of the PIMCity project is to make its data publicly available through public copyright licenses (e.g. Creative Commons), archiving it on the publicly accessible project website and OPENAIRE compliant repositories. Some of the PIMCity outputs are available for auditing, re-use and verification following the Open Access principles. Notice that the list of dataset results is limited to not make available dataset containing personal data, e.g., the data coming from the EasyPIMS main demonstrator and from the demonstrators involving the browsing history collected from the ISP DNS resolver and location history database.

5. Science Education

PIMCity partners recognise the importance of formal and informal science education in the society.[1] In relation to this, PIMCity partners contribute to it through several dissemination actions, e.g. by developing trainings and educational materials as revealed in detail below.

AUI, (WP6 leader) coordinates activities in order to contribute to science education. Overall, activities coordinated by AUI shall contribute to maximising the opportunities of adoption, increasing public acceptability and building new awareness and educational opportunities around personal data platforms. PIMCity partners prepared several tutorials on how to use the technology that has been developed during the project.

PDK training material has been made available through the https://easypims.pimcity-h2020.eu website. It contains both technical documentation (e.g., API description) and short videos to show how to use each tool. Videos are also available in the PIMCity channel on YouTube at https://www.youtube.com/channel/UC6rCkG6fj0UOQ0GpEbc_sqg.

In general, PIMCity training strategy addresses three different target groups: (I) potential end-users and stakeholders in the digital and data-centric businesses, (ii) society at large and (iii) University students.

According to this, PIMCity delivered online video tutorials providing guidance for the use of the tools developed for potential end-users and stakeholders in the digital and data-centric businesses, so that they would be able to educate themselves about the product use and configuration proactively. These video tutorials are freely accessible through the YouTube platform as well as from the project website.

PIMCity prepared 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. Some were imparted/translated into the local language of the audiences.

In detail, PIMCity partners prepared several training courses that were offered at various audiences. These include:

- Partners presented a short course on Human-Centered Machine Learning during the BIP Erasmus+ course at Aalto University Summer School and the Unite! Network. The course focused on the fair usage of data, and focused on the machine learning models and other algorithms that make use of all the data that is constantly generated. Human-Centered Machine Learning is an innovative and hands-on bachelor's level course on the main components and combinations of current machine learning systems. This two-week course will teach you some of the most widely used machine learning (ML) techniques. The focus will be on human-centered applications of ML methods that require high levels of personal data protection and transparency. The course includes lectures that teach basic principles of human-centered ML and its

---

[1] Science Education. European Commission: online access at https://ec.europa.eu/programmes/horizon2020/node/795 [accessed on 2020-11-11].

applications (such as elderly care). We presented how to implement privacy-preserving and transparent ML methods by using a few lines of Python during exercise sessions. These libraries included some of the tools of the PDK (the privacy preserving analytic tool). Details about the event can be found at https://www.pimcity-h2020.eu/event/human-centered-machine-learning/

- A workshop dedicated to high school students, aims to expand digital citizenship skills. The four lectures are available online at https://www.pimcity-h2020.eu/event/internet-pills-for-a-conscious-use-of-privacy/ The lectures are available online and are part of a workshop dedicated to professors and students from high school, covering topics related to the use of the Internet, such as i) Web tracking and targeted advertising, ii) GDPR and EU protection, iii) Data Monetization, iv) Data Control, v) Privacy and cybersecurity, Phishing, spearphishing, angler phishing and identity theft. The workshop will become part of the activities pre-Bienniale Tecnologia that will be promoted both as part of "Biennial for Schools". High school professors can join the workshop in presence at the Academy of Sciences.

- Partners presented a tutorial on the usage of the PDK at the MyData 2022 conference (May 2022) in Helsinky, in the format of a workshop titled "LEARN HOW TO IMPLEMENT A PIMS SOLUTION". In the workshop we showed how to implement a PIMS solution using the modules developed in the PIMCITY project, and shared the EasyPIMS experience and the lessons learned and results obtained. All these basic components offered by the PDK provide anyone the chance to build new PIMS and integrating modules in existing solutions. We describe how the Open APIs enable communications and interactions between components, easing integration of existing PIMS. The even can be found at https://2022.mydata.org/programme/

- Presented a tutorial at the Performance 2022 conference (November 2021) on "Privacy-Preserving Data Processing". In the tutorial, we presented the basic concepts of privacy-preserving data processing, introducing the main motivations, challenges and opportunities around anonymization techniques. We focussed on k-anonymity (and its variants) as the classical approach to anonymize tabular data, and we provided an overview of differential privacy as a flexible technique for running privacy-preserving queries. During the tutorial, the students put their hands on a real case-study and use Python code and state-of-the-art libraries to learn how to anonymize a dataset before publishing or querying it. All the material (slides, hands on material, data, and video recording of the lessons) can be accessed from the following link: https://www.performance2021.deib.polimi.it/tutorials/

Thirdly, academic partners will produce training materials for Master and Bachelor level courses. For example, UC3M, delivered training material that will become part of UC3M Masters courses (e.g., Big Data, Cybersecurity) and degrees (e.g., in Data Science), with particular attention to describe how the online ads ecosystem works. Similarly, POLITO has contributed to new courses in the Master's degree in ICT for Smart Societies (e.g., Big data for Internet applications, ICT for smart mobility) and Communications Engineering (Machine learning for networking, Network measurement laboratory), with the introduction on content about the privacy preserving analytics developed in the project.

Besides, PIMCity partners recognise that effective dissemination and communication contribute to science education. In relation to this, PIMCity partners assume that their dissemination and communication activities contribute to achieving their science education goals as well. As briefly mentioned in the part 1 of this document (public engagement), the reports of dissemination and communication are provided as deliverables of WP6 on the project's website pimcity-h2020.eu (see, e.g. D6.3 as completed on 30th November 2020 and D6.4 completed on 30 November 2021, and D6.5 completed on 31 August 2022).

6. Allocation of resources and data security

Making data FAIR has incurred no costs to the project. Datasets have been uploaded on Zenodo for free, which is a general purpose open-access repository under the European OpenAIRE program. In addition, by using Zenodo as the repository for project data set, we ensure their security and their long-term preservation.

Considering open access for publications, each institution has supported the eventual costs of required to make the articles appear as gold open access. For articles appearing as green open access, each partner has used the services offered by each institution, typically offered with no charges.

7. Ethical aspects

PIMCity partners recognise the importance of ethics as *an integral part of Research from the beginning to the end*.[2] PIMCity partners are aware that ethical research conduct *implies the application of fundamental ethical principles and legislation to scientific Research in all possible domains of Research,* including the domains researched in the light of PIMCity project. Taking into account that one of the most common ethical issues include privacy and data protection issues and the particular activities within the project, PIMCity partners have paid and pay particular attention to the compliance with all of the relevant national, European, EU and international privacy and data protection requirements. The KUL partner has monitored all the decisions related to data collection, storage and processing to advice on how the proposed solution can meet the strictest requirements of the European regulations. Besides, PIMCity partners aim to ensure there is no breach of research integrity, i.e. no falsification, plagiarism or other research misconduct.

PIMCity partners pay particular attention to the *European Commission Guidelines on ethics and data protection*.[3] Besides, since PIMCity partners worked on algorithms, the project pays particular attention to ethical requirements regarding the use of artificial intelligence. They take into account, among others, the *AI HLEG Ethics Guidelines for Trustworthy AI*[4] *and the Guidelines on Artificial Intelligence and Data Protection under the Council of Europe*.[5] Overall, the consortium works in line with these ethical values: (i) research should be designed, reviewed and undertaken to ensure integrity and quality; (ii) research staff and subjects must be informed fully about the purpose, methods and intended possible uses of the Research, what their participation in the Research entails and what risks, if any, are involved; (iii) the confidentiality of the information supplied by research subjects and the anonymity of respondents must be respected; (iv) research participants must participate in a voluntary way, free from any coercion, or risk; (v) harm to research participants must be avoided; (vi) the independence of Research must be clear, and any conflicts of interest or partiality must be explicit. The project partners will reveal their compliance with the ethical requirements through their individual inputs in detail.

Conclusions

This report presented the work performed by the PIMCITY consortium towards making the data created by the project FAIR. It summarizes the effort done by partners in making the results of the project easily accessible and reusable as much as possible.

---

[2] Ethics. European Commission: online access at https://ec.europa.eu/programmes/horizon2020/node/767 [accessed on 2020-11-11].
[3] Ethics and Data Protection. European Commission: online access at https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection_0.pdf [accessed on 2020-11-11].
[4] Ethics Guidelines for Trustworthy AI. European Commission: online access at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai [accessed on 2020-11-11].
[5] Guidelines on Artificial Intelligence and Data Protection. Council of Europe: online access at https://www.coe.int/en/web/artificial-intelligence/-/new-guidelines-on-artificial-intelligence-and-data-protection [accessed on 2020-11-11].