This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 953016

SERENDI PV

D7.2 Common data collection, QC and filtering, database and transfer protocols and standardization

T7.2 Common data collection, QC and filtering, database and transfer protocols and standardization.

Grant Agreement nº:	**953016**
Call:	H2020-LC-SC3-2020-RES-IA-CSA / LC-SC3-RES-33-2020
Project title:	**Smooth, REliable aNd Dispatchable Integration of PV in EU Grids**
Project acronym:	SERENDI-PV
Type of Action:	Innovation Action
Granted by:	European Climate, Infrastructure and Environment Executive Agency (CINEA)
Project coordinator:	Fundación TECNALIA Research & Innovation
Project website address:	*www.serendi-pv.eu; www.serendipv.eu*
Start date of the project:	October 2020
Duration:	48 months
Document Ref.:	D7.2 Common data collection, QC and filtering, database and transfer protocols and standardisation_v2.0.docx
Lead Beneficiary:	Mylight Systems (MLS)
Doc. Dissemination Level:	**PU** – Public
Due Date for Deliverable:	30/09/2024 (M48)
Actual Submission date:	17/10/2024 (M48)
Version	2.0 (Final Version)

Summary

This deliverable presents, through a public report, a set of long-term standard data protocols for use in data analyses and data exchanges in the photovoltaic domain. The standards cover the following themes: data collection, quality control and filtering, database storage and format, transfer protocols, standardizations and interoperability. A collaborative effort has been run within WP7, and in particular in this deliverable, to discuss with the stakeholders the standard of data sharing. The collaboration calls, organized within T7.5, have been used to gather feedback on the platform and the current sharing implementation.

Up until now, data collecting and sharing has been handled on a case by case "ad hoc" basis, which can cause problems at different stages of the lifetime of a dataset, be that due to misunderstood datetime formatting, wrong units on energy, incompatible storage formats, or poorly defined distribution rights. These standards have therefore been defined in order to facilitate the analysis and collection of data, as well as the exchange and sharing of data between collaborating partners. This report provides guidance to implement a data platform for renewable energy actors that need to perform analytics at scale on their data and share them with different stakeholders. By showing how SERENDI-PV partners tackle their business needs in terms of data storage, data exchange, data format… this report aims to help other actors to start and/or improve their platform. To confirm our recommendation, public surveys shared via social network (LinkedIn) and COPLASIMON contacts have been performed.

This deliverable is an output of task [7.2].

Document Information

Title	Common data collection, QC and filtering, database and format, transfer protocols, and standardisation and interoperability
Lead Beneficiary	Mylight Systems
Contributors	BI, QPV, LUC, SGIS, CYT, Akuo, MLS
Distribution	PU - Public
Report Name	D7.2 Common data collection, QC and filtering, database and format, transfer protocols, and standardisation and interoperability_v2.0

Document History

Date	Version	Prepared by	Organisation	Approved by	Notes
14/03/2023	Preliminary Version	J. Reed	MLS		
31/10/2023	Draft version	C. Salperwyck	MLS	Javier del Pozo (TEC)	Draft version (v1.0)
17/10/2024	Final version	C. Salperwyck	MLS	Javier del Pozo (TEC)	Final version (v2.0)

Acknowledgements

The work described in this publication has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 953016.

Disclaimer

This document reflects only the authors' view and not those of the European Commission. This work may rely on data from sources external to the members of the SERENDI-PV project Consortium. Members of the Consortium do not accept liability for loss or damage suffered by any third party as a result of errors or inaccuracies in such data. The information in this document is provided "as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and neither the European Commission nor any member of the SERENDI-PV Consortium is liable for any use that may be made of the information.

Contents

Tables

Figures

1 EXECUTIVE SUMMARY

1.1 Description of the deliverable content and purpose

This deliverable is a follow up from deliverable D1.4. which contains a set of recommendations and ideas within various, different "data themes" that the PV community should aim towards.

Within the photovoltaic domain there are a wide range of different datasets, data exchanges and use cases such as: production (or irradiance) prediction datasets for planning energy availability, balancing the electricity grid, monitoring of residential photovoltaic data energy production with the aim of optimising energy consumption… Those different use-cases and actors need to perform analytics and share data to accelerate the pace towards high-penetration of PV in Europe. Accompanying measures need to be set in place to lower the existing barriers to PV development, including the lack of common standards, protocols and good practices for data collection, exchange and storage.

This report focuses on that of a lack of a common set of data standards within the photovoltaic community. In lieu of a fixed set of rules for handling data, each case is handled in a sort of "ad-hoc" case by case way. This causes problems later, particularly during data exchanges. In this respect this report aims to act as a go-to guide with a set of good practices and recommendations.

In chapter 2, the focus is on gathering feedback on existing data platforms from SERENDI-PV partners and external actors. The covered topics are:

- Cloud vs on-premises to store and process data
- Which database to use (SQL, DataLake…)?
- Which granularity is useful?
- How to ensure and track data quality?
- How to exchange data and in which file format?

In chapter 3, the collaborative aspect through the COPLASIMON platform is presented. Different actions were conducted: collaboration calls, recommendation sharing at a conference, feedback, data privacy…

1.2 Reference material

The main documents taken into account when carried out this task T7.2, and used for the elaboration of this deliverable are:

- D1.3 Assessment of current and future grid financing challenges in a highly distributed power system and opportunities and threats to PV business models with high PV penetration.
- D1.4 Specifications on data collection, database, transfer protocols, data privacy and distribution and IP, which aimed to specify a set of long-term standard data protocols in the photovoltaic domain, as the work on this task T7.2 has taken as the starting point the preliminary work done in T1.4.
- D11.1 has been consulted for organizing the survey presented in chapter 2.

1.3 Relation with other activities in the project

Table 1.1 depicts the main links of this deliverable to other activities (work packages, tasks, deliverables, etc.) within SERENDI-PV project. The table should be considered along with the current document for further understanding of the deliverable contents and purpose.

Table 1.1: Relation between current deliverable and other activities in the project

Project activity	Relation with current deliverable
D1.3	This deliverable is a prospective analysis, assessing the financial challenges that could arise in case of a high penetration of PV in electricity grids, under various assumptions (e.g., PV penetration level, type of PV system considered, type of grid tariffs applied, ...). This deliverable also discussed the legal and ethical frameworks to share data that the COPLASIMON platform tackles.
D1.4	The aim of this task is to specify a set of long-term standard data protocols in the photovoltaic domain. The aim is to define, based on our expertise and experiences, a set of rules on data collection (what data), database storage and format (SQL/CSV, decimals/commas), transfer protocols (API/centralised platform), data privacy (accuracy of coordinates), distribution (rights to use data and for what) and Intellectual property (who owns the data).
T7.1	This deliverable presents the collaborative platform developed by SERENDI-PV, along with the current motivations behind this platform and the current status of advancement of the collaboration initiatives. This web-based platform has been named COllaborative PLAtform for SImulation and MONitoring (COPLASIMON), and it is currently hosted on a public domain: http://coplasimon.eu/
T10.2/ D10.3	D10.3 Data Management Plan details the plan for SERENDI-PV project to manage the data collected from the different PV systems and PV plants used to develop the project tasks and the research data generated during the project developments. It identifies the confidential datasets used, the public datasets available, and the scientific publications that will be related to confidential datasets (anonymizing the data).
T11.1 / D11.1	The WP 11 - Ethics requirements will address all the ethics issues identified in SERENDI-PV project. In particular, the current D11.1 H - Requirement No. 1 – Humans will address the Humans issues detected in SERENDI-PV. This deliverable (document) explicitly contains: • The procedures and criteria that will be used to identify/recruit research participants • Templates of the informed consent/assent forms and information sheets (in language and terms intelligible to the participants) • The procedure for the surveys: the invitation to participate, the design of the on-line anonymous survey and the gathering of the participants´ responses
T11.1 / D11.2	The WP 11 - Ethics requirements will address all the ethics issues identified in SERENDI-PV project. In particular, the current D11.2 Protection of Personal Data (POPD) - Requirement No. 2 will address the POPD procedures to be implemented in SERENDI-PV

1.4 Abbreviation list

Table 1.2: abbreviations used in the report

Abbreviation	Explanation
API	Application Programming Interface
AWS	Amazon Web Services
COPLASIMON	COllaborative PLAtform for SImulation and MONitoring
CSV	Comma-Separated Values
HTTPS	Hypertext Transfer Protocol Secure
JSON	JavaScript Object Notation
PV	Photovoltaic
SFTP	Secure File Transfer Protocol
SQL	Structured Query Language
TMY	Typical Meteorological Year
TSDB	Time-Series Database

2 FIELD FEEDBACK AND RECOMMENDATIONS ON ENERGY DATA STORAGE, PROCESSING AND SHARING

Monitoring PV systems is a fundamental part of the design, commissioning, and maintenance of all types of PV installations. This process includes data acquisition, data transmission, data storage and initial processing for a proper adequacy of the data.

There are several different objectives of PV system monitoring:

1. Evaluate (calculate) the performance of an individual PV plant,
2. Detect and identify faults and cause of underperformance,
3. Compare performances of systems (different configurations, locations…),
4. General monitoring (i.e., for residential installations)

The required data for monitoring are not the same depending on the objective:

1. For the first case, production data are needed as well as irradiation data to calculate standard metrics, such as the performance ratio, the array yield, etc. 10-minute data at plant level over a given duration may be sufficient. Various other metadata such as temperature, inclination, and orientation are also required.
2. For the second situation, a higher resolution is required with data at inverter or even string levels, with a recording interval of few tens of seconds or few minutes. More sensors of higher accuracy are also helpful to diagnose the origin of faults. Again, the irradiance and the various metadata are useful.
3. The third configuration may be in-between.
4. For simple monitoring of residential installations, just production may be required.

Deliverable D1.4 described current standards and practices as well as recommendations with regards to data collection, exchange and storage. In the following sections of this chapter, these initial recommendations are complemented by feedback from SERENDI PV partners based on their own experience with different aspects of data collection, exchange and storage as well as feedback from external stakeholders to the projects which were collected through a survey advertised through COPLASIMON. The survey covered questions focused on data size, data granularity, data databases, data exchange methods and data privacy. The exact questions included in the survey can be found in Section 5.2. The survey was disseminated through different channels such as emails leveraging some SERENDI-PV partners' network as well as posts on LinkedIn.

Five responders filled the survey. In addition to the survey answers, a bilateral exchange was organized with one of the responders allowing to gather additional information and feedback on the topic. Survey results are included in the subsequent sections and are exhaustively included in Section 5.2. As a general remark, the answers corroborate the initial recommendations provided in D1.4 thus reinforcing their relevancy for larger adoption in the PV community.

2.1 Cloud and on-premises partners data platform

Most of SERENDI-PV partners have Cloud-based platforms. Some of them also have on-site/on-premises machines/servers to collect and process locally data. This is especially needed for operations and maintenance of production and capacity sites.

On site storage allows high resilience of data: if issues arise with Internet connection, multiple backups are available to restore operations. Cloud services are a must-have in any modern data processing and storage platform. They allow fast and easy scalability and access to modern IT stacks at a reasonable cost.

Typically, data capture and storage from PV plants occupies a large amount of space, a local service has limited capacity and maintenance issues. Those problems are avoided in the Cloud, as it is easily scalable without major maintenance efforts. In addition, it allows rapid scalability and performance improvement for datasets that need it. It is also easier to have a Cloud SaaS solution for the main supervision and centralization of data.

In the case of SERENDI-PV partners, those Cloud services are mostly used for customer-oriented infrastructure. Customer-related data (such as project metadata, or third-party data from data loggers and meters) is stored in cloud databases and data lakes. The end-user applications and public API endpoints are also hosted in the cloud. It improves accessibility and speed for customers and ensure appropriately diversified backup of its data. Cloud services are used both for data storage and for the related services, improving the processing speed and reliability.

The results of the survey corroborate SERENDI-PV partners choices as nobody has only "on-premises" data platform but mostly in both locations as shown in Figure 2-1.

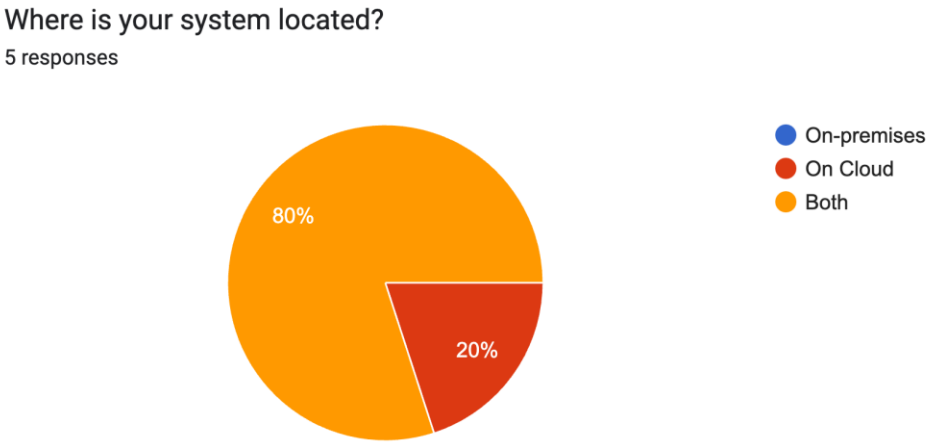


Figure 2-1 - External survey: data platform location

2.2 DataLake / Databases

There is a large variety of different options for storing data, depending on the quantity, the short- or long-term need for the data, the type of data, and the type of operations to be performed with the data. In this section, a variety of storage options will be discussed. Whilst data storage technologies are rapidly advancing, and any detailed long-term recommendations could quickly become obsolete.

A combination of on-premises and cloud storage seems to be very common. Scalability is the key factor since the database grows with the daily input of:

- energy data from customers
- satellite data
- energy prices from the market
- forecasting data
- …

Multiple types of storage options are used by partners:

- **Data lakes**
  - **On-premises**: some partners stored in Ceph (open-source, distributed storage) large and frequently ingested or updated data files (raw or processed solar, meteorological, and environmental multidimensional datasets in the NetCDF file format).

- **Cloud**: some partners use cloud storage as AWS S3 or Azure DataLake/BlobStorage (highly scalable and durable object storage in the cloud). For data reading and transformation, analytics tools like AWS Glue, Amazon Athena, Databricks, Kubernetes cluster or custom solutions (PAAS) are used. They allow rapid scaling and replicability.

- **Databases**: various SQL or non-SQL databases are used for structured data (tables, columns, rows) or document objects. Databases excel in fast querying and offer indexing and caching. Most small customer project's data and metadata are stored in databases.

For partners having plants with needs for operations, raw data are stored locally before being sent to the Cloud (partly or fully depending on the needs/volume) into cheap Cloud storage (such as S3 in AWS or BlobStorage/DataLake in Azure). This data tends to be large, bringing in large quantities of variables and with a low time periodicity. In order to have an easy and fast access on this plant data, some partners use a time-series databases (more details on such systems are presented in Section 3.3.1 of the public deliverable D1.4).

Databases are used to store results from data matching and analysis from the raw data. These are usually already focused on useful data, and their volume is therefore much less than the raw data. The use of databases allows a faster and simpler handling and can be used to set up your own, more controlled acquisition systems, such as an API service. Databases are also used for all “static” or “almost-static” data.

An important factor is the traceability of data within the database. Model changes and improves, over time it may produce different outputs for the same specifications, due to improvement in modelling or data sources. However, for customer projects and studies it is important to provide consistent data to ensure comparability and reliability. Therefore, it must be possible to find previous versions of models, code/algorithms and data in the database to explain any differences which happened over time.

The results of the survey corroborate SERENDI-PV partners choices as only once “Cloud Timeseries Database” was answered. Custom solutions with on-premises TSDB and SQL are preferred solutions, probably for cost reason and the ability to customize it for the specific needs. Surprisingly, software editor solutions (OSISoft, ABB, GE…) were not used by the people who responded to the survey as shown in Figure 2-2. Regarding the size of the data, Figure 2-3 shows that most use-cases need few terabytes of data. That explains the need/choice of Cloud base solutions to easily manipulate/compute this amount of data at scale.

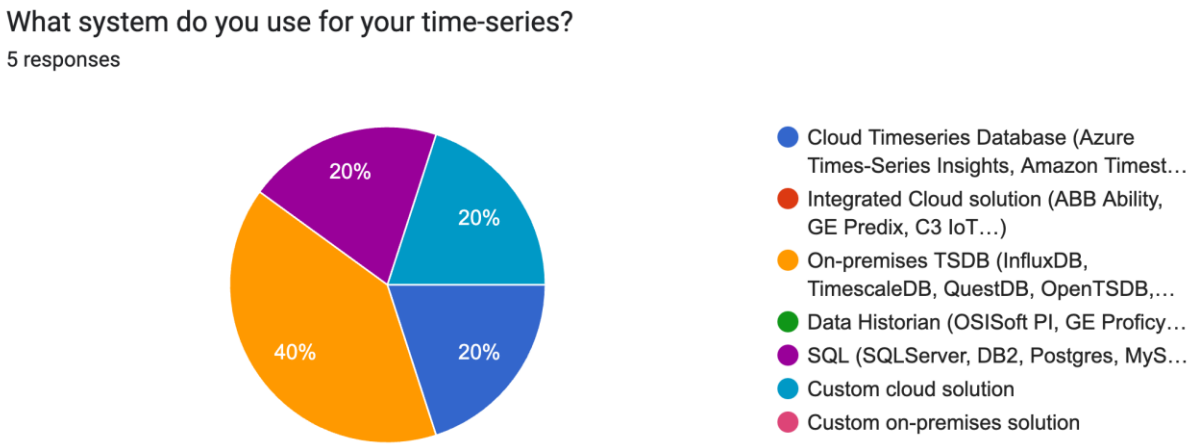


Figure 2-2 – External survey: time-series system used

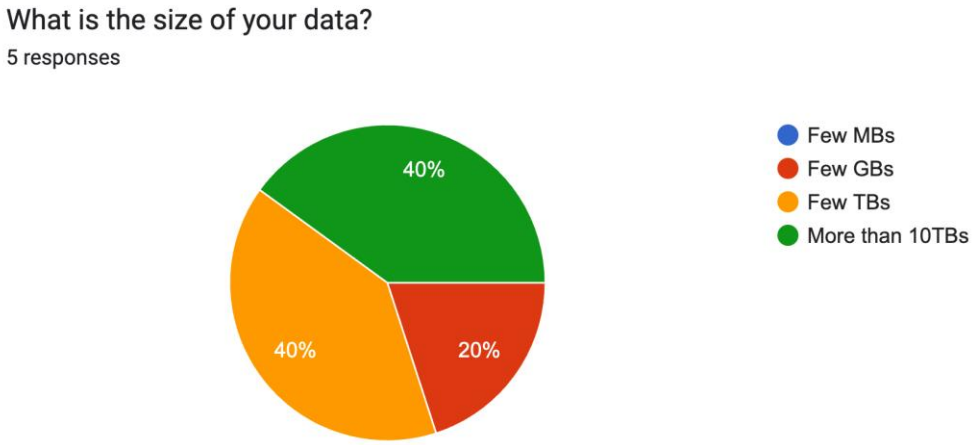


Figure 2-3 - External survey: data size

2.3 Granularity

Data often arrives with different granularity and synchrony. These data are generally stored as raw data and then processed. The cleaned and synchronized data is stored with a periodicity from 5 to 15 minutes in most of the use-cases of the partners, such as:

1. PV model to forecast PV energy for the next hours/days and/or PV simulation data to evaluate a PV project
2. Actual Energy data (PV production, site consumption) from customers, potentially linked with forecasted/actual irradiation/temperature
3. Performance analysis: from 15 minutes to longer time ranges (daily, weekly, monthly, quarterly or yearly).

Currently, the cost of storage and processing of a higher frequency does not provide more valuable results for those use-cases.

For on-site real time monitoring, the data are stored at the lowest granularity which is usually around 1 second. Usually in such systems to avoid overloading the system with data, only points that are sufficiently different from the previous value are stored (configured per measure).

Figure 2-4 shows the results of the survey and confirm the granularity to be mostly “from 5 to 15 minutes” for Cloud base system, and from “1s to 60s” for on-premises systems. Note: using individual survey we can match granularity answers with Cloud/on-premises ones.

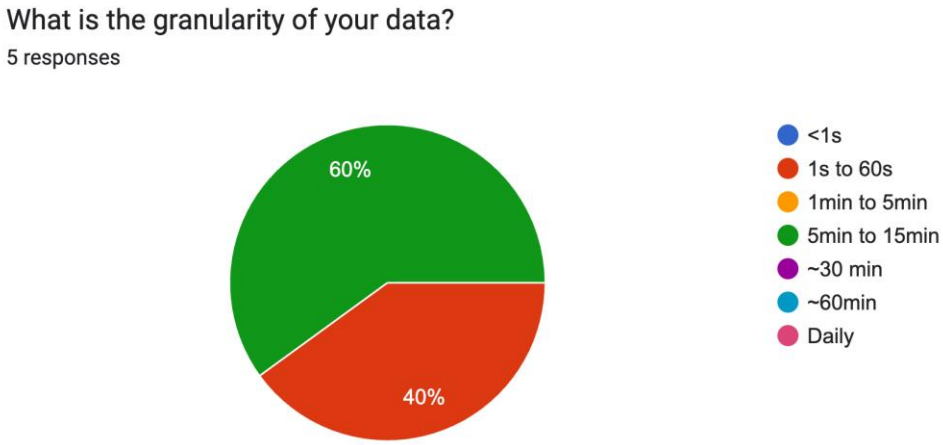


Figure 2-4 - External survey: granularity of data

2.4 Quality

Based on data from the customer (such as location or PV plant specifications), the solar and meteorological model, and the PV yield simulation are appropriately configured. Furthermore, the desired time granularity is selected (although all data is natively produced with 15-minute granularity).

When data are shared, it must include the metadata needed for using / understanding the data:

- Customer specifics such as the location, kWp installed…
- Source and resolution of the various groups of data (solar, meteorological, hardware on the customer site…)
- Commercial information (when it comes from a product)
- Description of the parameters within the file, including units, and any invalid data flags used (e.g. using "-9" for missing values, number of missing days of data for yearly aggregated data…).

Quality and accuracy of the provided data is also an important consideration. Partners employs several methods to ensure acceptable accuracy and quality, among them:

- Model and simulation data are compared to real-world measured data to evaluate the bias and error in the models. These validations provide an understanding of the model performance in various locations and climates around the world.
- Data sources are updated to take advantage of the improvements in technology - e.g. data from new satellites are provided with greater temporal and spatial resolution compared to their predecessors.
- The underlying solar, meteorological, and PV yield models are constantly evaluated and improved to decrease the observed errors compared to ground measured data and provide new features (such as snow and soiling loss models, developed within SERENDI-PV task T2.2).
- Time consistency: data covering the entire time span, with no gaps in between. In addition, they must be synchronous with the entire plant.
- Data filtering: these filters follow IEC-61724 standards, which ensure the validity of the data.
- Advanced filtering: patterns that are not possible for physical or operational reasons are located. This is done by crossing data from multiple devices.

- Quality attributes/tags on data at each step: SCADA/source, monitoring, performance analysis (with tools such as PowerBI for standard reporting or Bazefield for more advanced reports)

Furthermore, partners developed tools to validate the data provided by customers as inputs, as its accuracy is key to ensure provision of the right model/simulation. Within this scope there are algorithms which analyse ground-measured PV data and detect real PV system configuration, identify plant faults, and snow and soiling impact on power production. These algorithms were developed within the scope of Work Package 3 of the SERENDI-PV project. Partners also employ processes for quality checks of customer data to ensure that systematic issues and failures (e.g. instrument malfunction, cloudy weather, inverted measurements…) are excluded from the data analysed. In this manner the further processing of data is not affected by these errors and the results (e.g. model or simulated data) are more accurate.

2.5 Data exchange format

In the following section we explain which format are used to exchange data between

- customers and partners
- between partners

There are two main categories of format that were used during the SERENDI-PV project:

- Human-readable:
  - comma-separated values (CSV) format for larger volume: it enables the customer to unambiguously interpret the data provided. REST API call can return a link for direct download of the CSV file.
  - JSON for smaller volume and synchronous REST API calls
  - standardized JSON for metadata: the data definition is published online as JSON schema documents. Two main data types are based on these definitions - datasets and requests. Standardized datasets are typically provided to users, but can also be sent by users e.g., when sharing data from data loggers or various instruments. Requests are used as API payloads.
- Binary format:
  - Parquet file: for sharing large volume of energy data (PV production, load curves) in an efficient manner. Those files can be compressed using the efficient zstd library - size of the file can be divided by a factor of 5 to 10, leading to storage savings. This format has the advantage of being a columnar storage, avoiding reading unnecessary data when not requested. It also stores the types avoiding type conversion that are prone to errors.

The results of the survey in Figure 2-5 shows that human readable format is preferred to binary/optimized format.

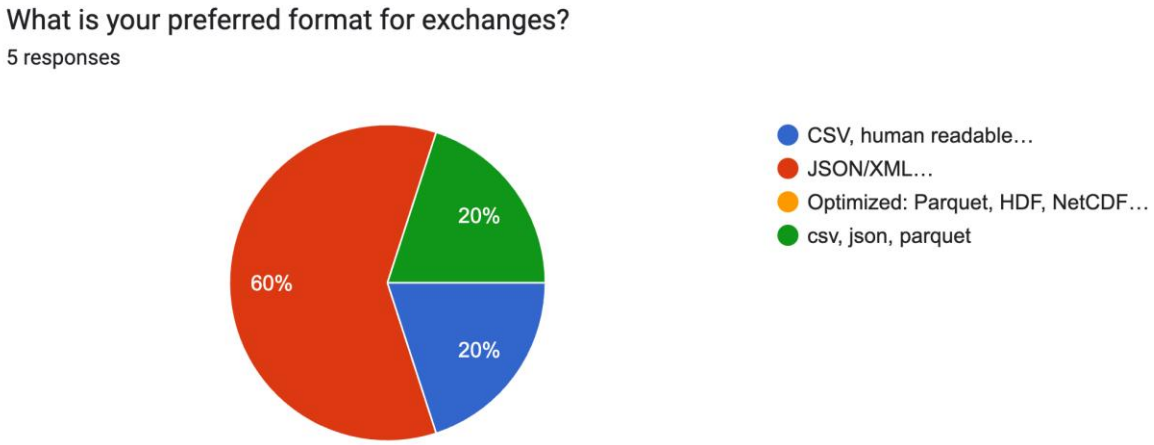


Figure 2-5 - External survey: preferred exchange format

2.6 Transfer Protocols

Data needs to be shared, between collaborating parties in a research context, or between electric grid management services that need to know how much energy is being injected where and when, or between grid management companies and electricity companies wanting to charge clients. Data transfer solutions and protocols are needed to enable simple, fast and reliable, exchange of data. The type of transfer depends on several factors including the quantity of data to be transferred, its format, the required security, and speed.

2.6.1 API for customers

In general, exchanges are achieved by using Application Programming Interface (API), both for providing and receiving data. APIs can be connected to automated services to speed up the necessary data processing. In this manner the data exchange is greatly simplified and allows any partner to connect without significant effort. Modern IT solutions for data exchange are standardized and properly documented using tools as Swagger.

Most Cloud providers provide API endpoints that can scale according to load and can be called efficiently in a parallel mode. Recent API endpoints use standardized JSON data structures in request or response payloads. In case the data are on-premises, Secure File Transfer Protocol (SFTP) from the server is the recommended options that partners use.

Synchronous API operations are used for smaller datasets such as:

- monitoring of recent history
- forecasting
- listing and updates of metadata
- results of analysis
- …

Asynchronous mode is used for getting bigger datasets - multi-year timeseries or TMYs.

API endpoints use HTTPS transfer protocol ensuring that sensitive payloads transmitted between the client and server are encrypted.

Connection of actual customer data with simulation/predicted data is optimal to assess/correct the models. For that reason, a platform supporting bi-directional communication is the goal of the development in data

exchange services. This platform must respect the principles of confidentiality, copyrights attribution, and ownership, as it contains proprietary customer data. For that reason, the platform is a closed system.

Partners also participate in open sharing of their data for research purposes. Data in the form of maps and map layers (suitable for Geographic Information Systems - GIS) are published. Studies of solar PV potential of countries around the world with the accompanying data relevant to PV power are published as a part of the Global Solar Atlas. As an example, Solargis Prospect, one of Solargis' core products, is based on a freemium model where several data layers in the form of maps can be explored for free. Solargis also provides its model data to students and research institutions.

2.6.2 Limited time data sharing

In case a large amount of data needs to be shared (in the order of hundreds of megabytes) for a one-time usage/study, developing a specific API would require too much effort. Cloud services offer temporary data sharing using a "temporary key" that expires after a given date. It allows to share data for only few hours/days to only a set of people to avoid exposing those data later in case the "key" get stolen. It is usually a good practice to share the credentials using 2 different emails: one with the link/URL and one with the key. If an email is forwarded by mistake, the recipient has only just a part of the information and cannot access the data. You can also configure what permissions the shared resource has: read only, write, append… In case the data are on-premises, Secure File Transfer Protocol (SFTP) from the server is the recommended options that partners use.

The results of the survey corroborate SERENDI-PV partners choices in Figure 2-6: API and FTP are the most used. Queuing systems (mentioned in D1.4) also appear for low granularity use-cases for which they are particularly suitable.

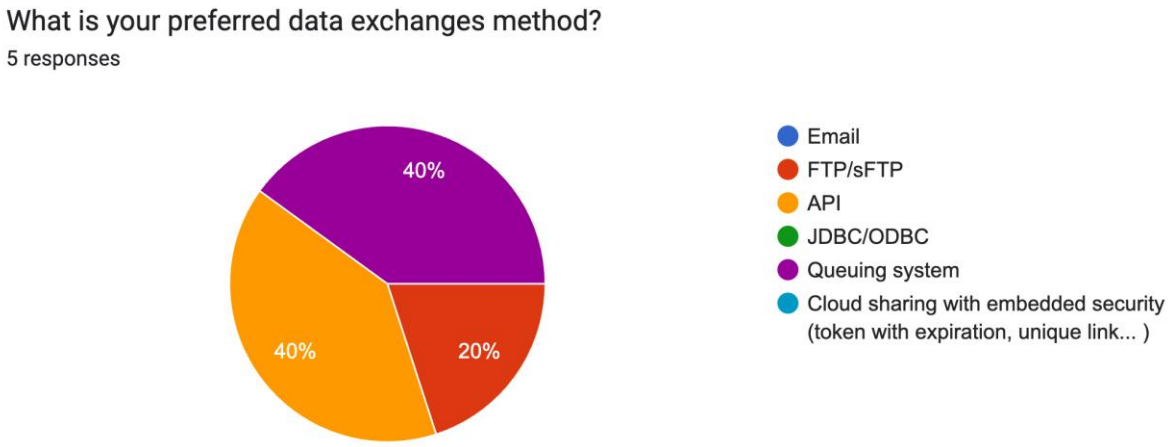


Figure 2-6 - External survey: data exchanges

2.7 Survey: exchange with one of the companies

One of the companies that replied to the survey gave us their contact details. A call was scheduled with them in September 2024 to get more details about their use-cases and data platform. The interview followed the same questions as the ones in the survey.

This company develops a new product targeted to wind and PV industry to collect data and perform actions in real time in their farm. There is an important need to be able to connect assets to energy demand such as negative prices, system services...

The useful data granularity is in minutes for now, but they plan to lower it to seconds to be able to take decisions in only few seconds to address some use-cases related to energy demand.

They use TimescaleDB on a private datacentre as their customers do not feel at ease using the public clouds.

The communication between the farms and the datacentre might be chaotic, therefore 4G/5G networks or satellite communications are welcome.

The industry is looking for cost efficient IT solutions to analyse their large amount of data. There is a need to access large amount of data to perform data analysis/science that standard monitoring tools do not allow as they are not designed for massive data extraction.

The deliverable D1.4 was mentioned during the interview and our recommendations matched their findings to build their data platform. The deliverable D1.4 and the discussion around the contents of deliverable D7.2 helps them to keep going with their current platform without challenging their choices.

3 COLLABORATIVE PLATFORM: FIELD FEEDBACK, RECOMMENDATIONS, AND IMPLEMENTATION

3.1 Introduction to the COPLASIMON platform

COPLASIMON is the collaborative platform of the SERENDI-PV project.

This platform goal is to test the project innovations on real cases giving feedback to the PV community.

This collaborative platform constitutes a vehicle for collaboration between the partners of SERENDI-PV and the other relevant stakeholders from the solar energy community.

This platform deals with topics such as simulation, design, data analytics, QC and grid integration. The platform content includes the main developments of the SERENDI-PV project and include:

- State of the art of PV and of the SERENDI-PV innovations
- Collaboration calls to test the software developed in the project on real use cases
- Organization of workshops (e.g. the soiling workshop organized within T7.7) on several topics and interaction with stakeholders to understand their needs
- Platform for code sharing in Quality Check and data analysis (GitHub).
- Publication of open-source tools or DEMO versions through link to the platform
- Storage of public data on PV shared by external stakeholders and SERENDI-PV partners (CKAN database)
- Forum for discussions among the PV community

LuciSun has developed, in the scope of WP7, collaborations and data exchanges with external partners to the project to test the software and the innovations of the project.

These software tools and innovations require validation using experimental data from stakeholders. That is why COPLASIMON opened collaboration calls to open the SERENDI-PV tools to the external stakeholders.

Several tools which have been put in place to make collaboration easier when required, such as:

- **A forum for discussion**. The forum has been built upon the **Discourse** framework because it is widely used for similar collaboration projects and because of its flexibility.
- **A repository for source code exchange** and developments including a versioning system. **GitHub** has been used for that purpose, differing from the starting page in Gitlab, following changes in licensing rules on Gitlab's side.
- **A database** for the structured organization of the information. **CKAN** has been selected for the database, because it is the most commonly used in data science, and because of previous successful uses in other European research projects on PV monitoring and performance, such as the COST project Pearl PV.

The inputs from the D1.4 have been used during the data exchange with the aim to implement different protocols in the data exchange and have guided the redaction of the surveys available for the public and described in chapter 5.

A documentation has also been produced for the COPLASIMON users allowing to share data on three different ways:

- Open access for the public

- Open access to the SERENDI-PV partners and COPLASIMON board members
- Private access for dedicated partners

The access can be granted in these different ways in COPLASIMON and there is the possibility to dynamically change the access privileges.

Among the WP7 several tasks require the data sharing between partners and among the industrial and research context. The type of transfer depends on several factors including the quantity of data to be transferred, its format, the required security, and speed. All these elements were taken into account while developing the file transfer and APIs (JSON, CSV, TXT, parquet format).

Collaboration calls have been opened on COPLASIMON (http://coplasimon.eu) and it has been decided to have them permanently opened in order to grant a continuity of the project and a transition toward an autonomous platform able to self-sustain and increase the number of users.

The calls cover the following SERENDI-PV project topics:

- Bifacial PV
- Floating PV
- Soiling
- PV Degradation
- Financial Tools for PV
- BIPV

For the calls abovementioned the SERENDI-PV partners participating offer to the data providers:

- A free QC of the data
- Data analytics
- Reports with the main results and recommendations

Within task T7.2 and the collaboration calls (T7.5), LuciSun has gathered feedback on how the COPLASIMON platform should adapt depending on the data required and proposed. The aim is to see which is the best data exchange methodology, depending on the topic to treat, the minimal requirements and data availability.

The discussion on this topic has been fuelled by the different discussions in the forum and by email in which the partners have interacted with the PV community on the needs for the data exchange and have engaged discussions with the data providers in the framework of the collaboration calls. One example of this collaboration is provided in section 3.2.

3.2 Collaborations with other initiatives on PV digitalisation (IEA PVPS and ETIP PV)

LuciSun, in the context of COPLASIMON, has been collaborating with other initiatives on the topic of PV digitalization. Two collaborations have been particularly active over the last years:

- Collaboration with IEA PVPS Task 13
- Collaboration with ETIP PV

Joint effort has been developed as part of a collaboration between COPLASIMON and ETIP PV to consolidate efforts towards the digitalisation of the PV sector. Digitalisation is a pivotal issue that influences every segment of the PV value chain, impacting not only energy production from PV technologies but also the integration of renewable resources into the broader power system. Proper digitalisation is essential to monitor, control, and ensure the stability of electricity systems. However, digitalisation is a multifaceted concept with varied implications, benefits, and challenges depending on the specific stage of the PV value chain.

The results of this work have been presented in a plenary oral conference at EU PVSEC 2023, Lisbon[1]. Drafted within the framework of ETIP PV's Working Group on Digital PV and Grid, this paper aims to provide a comprehensive analysis of the role and impact of digitalisation in the PV industry. It offers a detailed mapping of the key technologies, benefits, and risks associated with digitalisation, and addresses the barriers and potential best practices to fully harness digital tools in the sector. The collaboration between COPLASIMON and ETIP PV focuses on establishing a shared foundation to maximise the potential of digital solutions, facilitating the implementation of best practices across the sector.

This work delves into several critical aspects of digitalisation for PV, beginning with its contribution to the manufacturing process and dataflows within the industry, which are essential for delivering high-quality installations at a lower cost and with better environmental performance. It further explores advanced digitalisation pathways, such as interfacing PV data fusion with digital twins, integrating BIM and GIS data, and using hyperspectral and multispectral imagery for system optimisation and quality assurance.

Additionally, the study addresses the integration of digital PV within energy communities, highlighting its role as a tool for flexibility, improved forecasting, and enhanced grid impact management. As digitalisation becomes increasingly embedded in energy systems, cybersecurity and data integrity emerge as paramount concerns. This collaboration underscores the need for developing robust best practices to safeguard digital infrastructures, especially in a context where digital and energy sectors are becoming more interdependent.

Moreover, the paper analyses policy and regulatory challenges associated with digitalisation in the electricity market, with a specific focus on data-sharing frameworks in PV tenders and the digital considerations embedded in the NECPs. By mutualising efforts and pooling expertise, this joint initiative between COPLASIMON and ETIP PV aims to establish a clear roadmap for digital transformation, creating a unified approach that benefits all stakeholders across the PV value chain.

Collaboration with IEA PVPS has included PV digitalisation, in particular regarding PV monitoring data, PV dataflow for modelling, digital twins, and FAIRification, and standardisation. The first results from the collaboration on PV digitalisation with IEA PVPS, which is still ongoing, was presented at Intersolar 2024 in an oral conference titled, "Mapping the Relevance of Digitalization for Photovoltaics" [2], delivered by Jonathan Leloux, outlined the significant strides made in the digitalization of the PV sector, in line with the goals of the COPLASIMON initiative and the outcomes of the SERENDI-PV project. It was highlighted that the rapid expansion of photovoltaic projects had outpaced the availability of experts and the evolution of products, making digital solutions essential. The SERENDI-PV project, funded by the European Commission through Horizon 2020, played a key role in advancing these digital technologies to meet these demands.

During the presentation, the challenges within the PV simulation chain were addressed. It was demonstrated that the fragmented processes and lack of integration had led to inefficiencies. Platforms such as PVsyst and PVcase had been introduced to improve data flow between users, though there remained considerable room for further digitalization. COPLASIMON has provided a shared ecosystem where these gaps can be bridged, bringing stakeholders together to create a more cohesive and reliable PV data network.

Several significant public data-sharing initiatives, including IEA PVPS Task 13 [3] and the PV fleet performance data project by NREL [4], were highlighted as important steps toward improving transparency and reliability in PV performance data. These initiatives have been closely linked with COPLASIMON's mission, as better access to data and collaboration were essential in closing the gap between field performance data and simulations.

The application of PV performance diagnostics, as part of the SERENDI-PV project, have been a significant milestone in automating and accelerating PV performance reports and diagnosing system issues more rapidly. Digital twins have been underlined as a potential key training method, allowing professionals to simulate maintenance tasks and enhance system reliability without risking real-world operations.

Stakeholders within COPLASIMON have been invited to engage with these results, particularly regarding how the digital tools developed by SERENDI-PV—such as digital twins and data-sharing platforms—could be further integrated into their operations. Feedback was actively sought to refine these innovations and ensure

they met the practical needs of the PV sector. This collaborative effort has been essential in shaping the future of digitalization in photovoltaics, ensuring the solutions developed were practical, scalable, and widely adoptable within the industry.

The achievements presented underscored the importance of continuing to optimize PV system digitalization, paving the way for a more resilient and efficient solar energy infrastructure.

3.3 Feedback from the collaboration calls

The collaboration calls have been successfully led by the project partners, with several participants joining the initiatives.

As a follow-up to collaboration activities conducted through the COPLASIMON collaborative platform, feedback was collected from COPLASIMON users (both from the data provider and the data receiver) in order to highlight potential areas of improvements for future data sharing activities. The survey covered questions focused on data type, data granularity, data cleaning and treatment, formats of datasets, exchange/transfer protocols and data privacy. The exact questions included in the survey can be found in Section 5.1. The survey was disseminated to the COPLASIMON users (or stakeholders which had expressed their interest in using the COPLASIMON platform) and the dissemination is ongoing.

The survey is a long-lasting initiative which will outlive the project and will remain implemented as part of the COPLASIMON platform after project end and, in line with the objectives of the COPLASIMON platform, the survey is intended to be conducted after each data exchange taking place on the platform even beyond SERENDI-PV project end. Eventually, results collected after the project end will contribute to the improvement of data collaboration taking place on the platform.

Both data providers and researchers expressed significant interest in joining the collaboration calls. However, once the collaboration begins, the lack of funding often reduces the number of active collaborations which are led until the end. LuciSun has led several collaborations through COPLASIMON and most of them have led to scientific publications. However, the funding of the SERENDI-PV project and COPLASIMON has demonstrated as a key element for sustaining these collaborations. It is recommended that appropriate funding be allocated to strengthen the collaboration between laboratories, academia, and industry to encourage more sustained partnerships. In USA there have been calls (i.e.: “Solar Data Bounty Prize” [5]) for data on the PV which have been funded by the government in order to attract the data providers (private companies) or prosumers (for residential installations).

Another major issue identified relates to the availability of data from providers. Data providers typically need to invest time to clean and prepare the data for public sharing, respond to questions, and review the results. Without additional funding, it is challenging for private companies, which usually manage the data, to share it in a format and condition suitable for researchers. On the other hand, if researchers find that the data is not in the correct format, is incomplete, contains missing values, or lacks clear metadata, they tend to discontinue or abandon the collaboration.

The lack of accessible and properly formatted data has been one of the most significant obstacles in advancing specific collaborations within COPLASIMON as the survey revealed. The most commonly used format is CSV due to its ease of visualization. However, the NetCDF format may be preferred for handling larger datasets.

All these obstacles can be overcome by putting in contact stakeholders who have already a funding but are looking to extend their network and prepare future projects.

The partners of the SERENDI-PV project are encouraged, through conferences ([6]) and their network to promote new stakeholders to join the platform and participate in future collaboration calls via COPLASIMON.

3.4 Data Privacy, Sovereignty, and Ownership

This section covers several data themes that overlap with each other, that of data privacy, sovereignty, security and rights (i.e., who owns and has the right to do what with what). Data privacy and sovereignty laws and the rules that govern the handling and sharing of both personal and confidential data are constantly evolving and changing. With the arrival of the internet and the digitalisation and delocalisation of data (as in any sort of data can now be transferred, exchanged, bought, sold etc), the laws have had to evolve to protect individual privacy, security and intellectual property considerations. The international-ness of data transfers and the internet in general means that, sometimes, differing rules, customs and laws can come up against each other. Large political events like Brexit (UK exit from GDPR), or Social Media company monopolies (Meta's threat to leave Europe following stricter data privacy laws), or private data leaks (too numerous to list) can all effect the landscape of data regulation.

This deliverable, compared to D7.3 and D1.4, has tackled during the collaborations which are the best measures that can be of real application in an international collaboration involving several partners.

D1.3 has suggested that there is a "*lack of robust legal and ethical frameworks, as well as governance models and trusted intermediaries that guarantee data quality, reliability, and its fair use".*

This is why the COPLASIMON organization will let the partners interact between them providing standard templates for the agreement but let the freedom to the partners for tackling the specifies of the data sharing. The main legal requirement for the COPLASIMON webpage will be to treat the request of contact and collaboration with the partners.
It means that, after publishing one collaboration call, the partner is invited to fill in a document with the personal data and the specifications of why they participate to the collaboration. This process allows to draft the agreements that are signed by the individual parties.

The platform followed, in particular, the guidelines and requirements while working with data and external users defined in D10.3, D11.1 and D11.2.

3.5 Data Anonymization Specification for PV Data

This section details the methods and protocols to be employed to ensure the photovoltaic (PV) data used in COPLASIMON is anonymized. This process is vital for safeguarding specific site details, such as location, while preserving the general utility of the data. The primary objective of this data anonymization process is to prevent the identification of individual PV installation sites and other potentially identifiable data, ensuring data security and relevance for our project purposes.

Several methods have been identified and, following the discussion with the external partners, they are generally applied on the data:

- **Location obfuscation**: To protect the identification of specific PV installation sites, the precise location data will either be excluded or generalized to a lower resolution (e.g., city or district level rather than exact coordinates).
- **De-identification**: All direct identifiers that might point to a particular installation or entity will be removed.
- **Pseudonymization**: Original identifiers will be replaced with artificial ones, ensuring that backtracking to the original data is not feasible.
- **Aggregation**: Data will be summarized in such a way that individual data points related to specific PV installations cannot be isolated.
- **Data masking:** Techniques like data shuffling or substitution will be used to obscure specific data within the dataset.

- **Differential privacy**: Random noise might be introduced to the data to protect the identification of particular PV installations while retaining the overall trends and patterns in the data.

By adopting the methods described above, we ensure that the PV data used in COPLASIMON remains relevant for analysis without compromising the privacy and security of specific installation sites.

4 WIKIPEDIA PAGES CONTRIBUTIONS

The SERENDI-PV partners have decided to interact with projects and resources external to the project and, with this purpose, contribute to some Wikipedia pages.

4.1 Pyranometer

A section was added in the "usage" part of the Wikipedia page (Figure 4-1) related to Quality Assessment of data measured by pyranometers:

https://en.wikipedia.org/wiki/Pyranometer#Usage

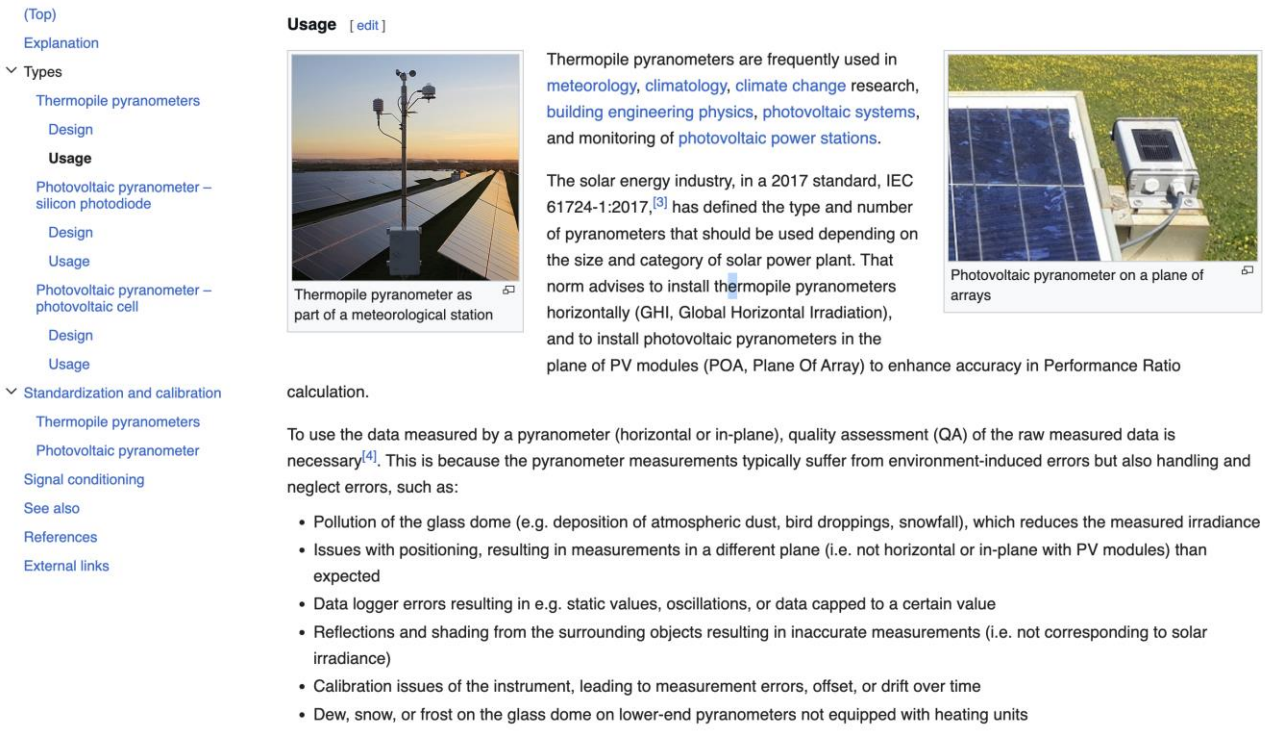


Figure 4-1 - Wikipedia Pyranometer page with project contribution

This section describes "the typical issues of raw data measured by thermopile pyranometer, and the quality assessment procedure that is typically required for the raw data" – the contribution is the following:

To use the data measured by a pyranometer (horizontal or in-plane), quality assessment (QA) of the raw measured data is necessary. This is because the pyranometer measurements typically suffer from environment-induced errors but also handling and neglect errors, such as:

- *Pollution of the glass dome (e.g. deposition of atmospheric dust, bird droppings, snowfall), which reduces the measured irradiance*
- *Issues with positioning, resulting in measurements in a different plane (i.e. not horizontal or in-plane with PV modules) than expected*
- *Data logger errors resulting in e.g. static values, oscillations, or data capped to a certain value*
- *Reflections and shading from the surrounding objects resulting in inaccurate measurements (i.e. not corresponding to solar irradiance)*
- *Calibration issues of the instrument, leading to measurement errors, offset, or drift over time*
- *Dew, snow, or frost on the glass dome on lower-end pyranometers not equipped with heating units*

Each of the above issues appears as a specific pattern in the measured time series. Thanks to this, the issues can be identified, the erroneous records flagged, and excluded from the dataset. The methods employed for data QA can be either manual, relying on an expert to identify the patterns, or automated, where an algorithm does the job. As many of the patterns are complex, not easily described, and require a particular context, manual QA is very common. A specialist software with suitable tools is required to perform the QA.

After the QA procedure, the remaining 'clean' dataset reflects the solar irradiance at the measurement site to within the uncertainty of measurement of the instrument. The 'clean' measured dataset can be optionally enhanced with data from a satellite-based solar irradiance model. This data is available globally for a much longer time period (typically decades into the past) than the data measured by the pyranometer. The satellite model data can be correlated (or site adapted) to the pyranometer-measured data to produce a dataset with a long time period of data accurate for the specific site, with a defined uncertainty. Such data can be used to perform bankable solar resource studies or produce Solar potential maps.

For monitoring of operational PV power plants, pyranometers play an essential role in verifying the solar irradiance available at any given time or over a certain time period. Due to weather variability, redundancy, and the spatial scale of contemporary solar plants (above 100MWp), multiple pyranometers are installed to provide accurate solar irradiation for each section of the PV power plant. IEC 61724-1:2017 international standard for example calls for at least 4 Class A thermopile pyranometers to be installed at 100MWp PV power plant at all times.

Solar measurements that were QA'd could be used to derive Key Performance Indicators (KPI) such as Performance ratio - metrics used in asset health monitoring or various contractual scenarios relating to energy produced (billing) or asset management (i.e. O&M). In these calculations, the measured sum of in-plane irradiation over a certain period is used as the determinant to which normalized produced PV electricity is compared to. Due to the difficulty of obtaining reliable in-plane measurements, especially in operational power plants, Energy Performance Index is increasingly being used instead of the older Performance ratio metric.

4.2 Industrial Internet of Things

A section was also added in the "Application and Industries" part of the Wikipedia page on IIoT (Figure 4-2):

https://en.wikipedia.org/wiki/Industrial_internet_of_things

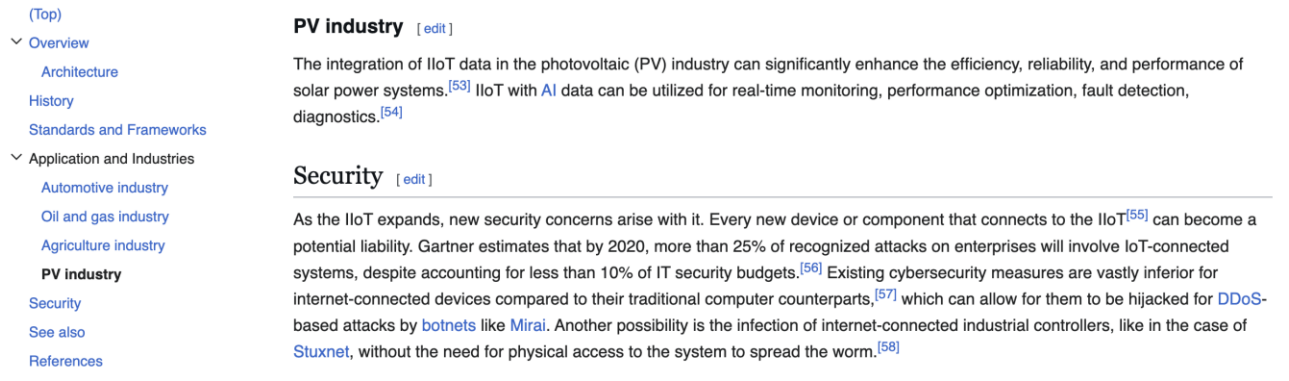


Figure 4-2 - Wikipedia IIoT page with project contribution

The contribution was the following:

The integration of IIoT data in the photovoltaic (PV) industry can significantly enhance the efficiency, reliability, and performance of solar power systems. IIoT with AI data can be utilized for real-time monitoring, performance optimization, fault detection, diagnostics.

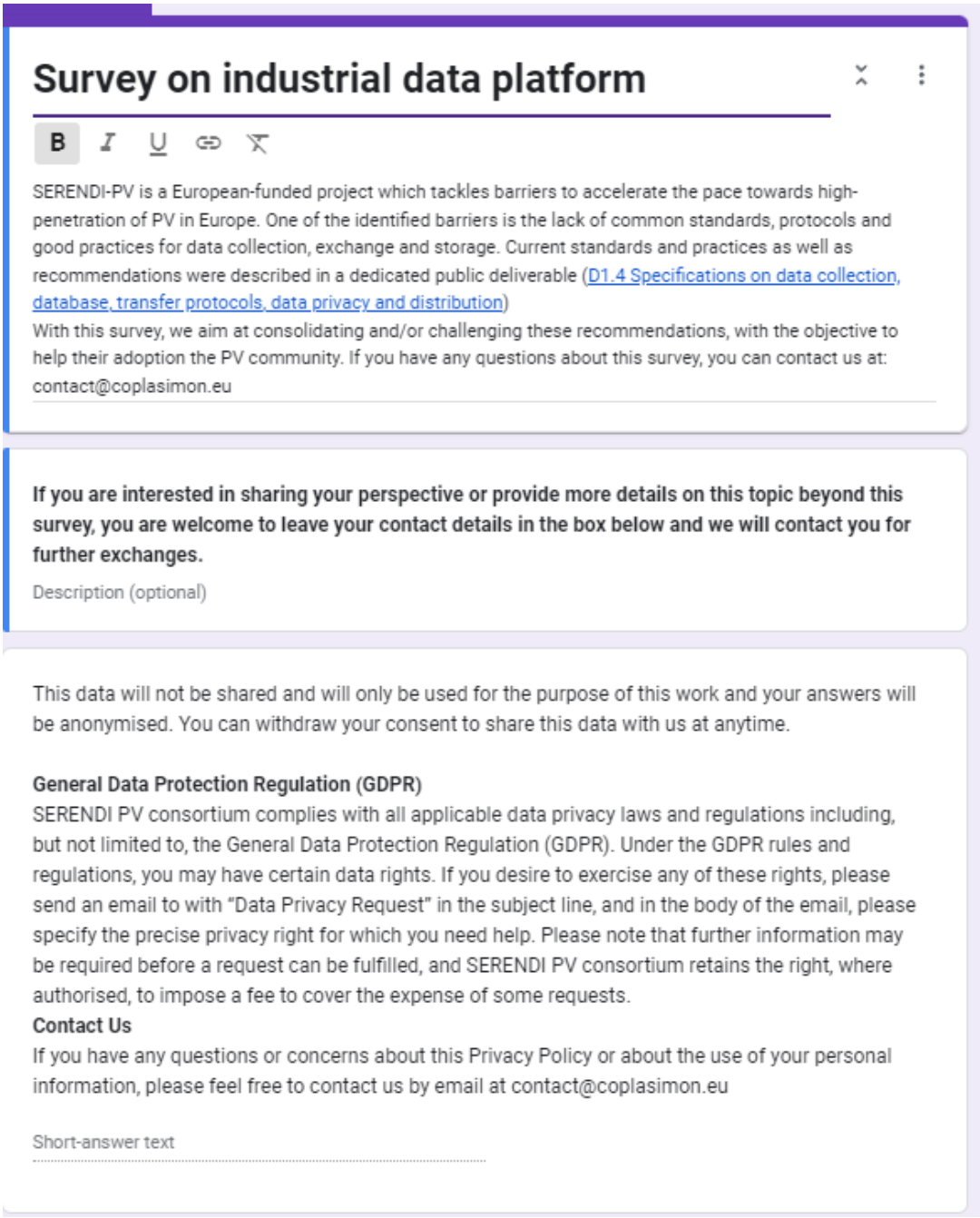
We also did a minor contribution on the section “AI/Machine Learning”: *There are many use-cases using AI with IIoT, to name a few: condition monitoring and predictive maintenance, process optimization, federate learning…*

5 APPENDIXES

5.1 Survey on data collection, exchange and storage oriented towards COPLASIMON users.

The survey is constructed as follow

- **Introduction and GDPR aspects**



Survey on industrial data platform

SERENDI-PV is a European-funded project which tackles barriers to accelerate the pace towards high-penetration of PV in Europe. One of the identified barriers is the lack of common standards, protocols and good practices for data collection, exchange and storage. Current standards and practices as well as recommendations were described in a dedicated public deliverable (D1.4 Specifications on data collection, database, transfer protocols, data privacy and distribution)
With this survey, we aim at consolidating and/or challenging these recommendations, with the objective to help their adoption the PV community. If you have any questions about this survey, you can contact us at: contact@coplasimon.eu

If you are interested in sharing your perspective or provide more details on this topic beyond this survey, you are welcome to leave your contact details in the box below and we will contact you for further exchanges.

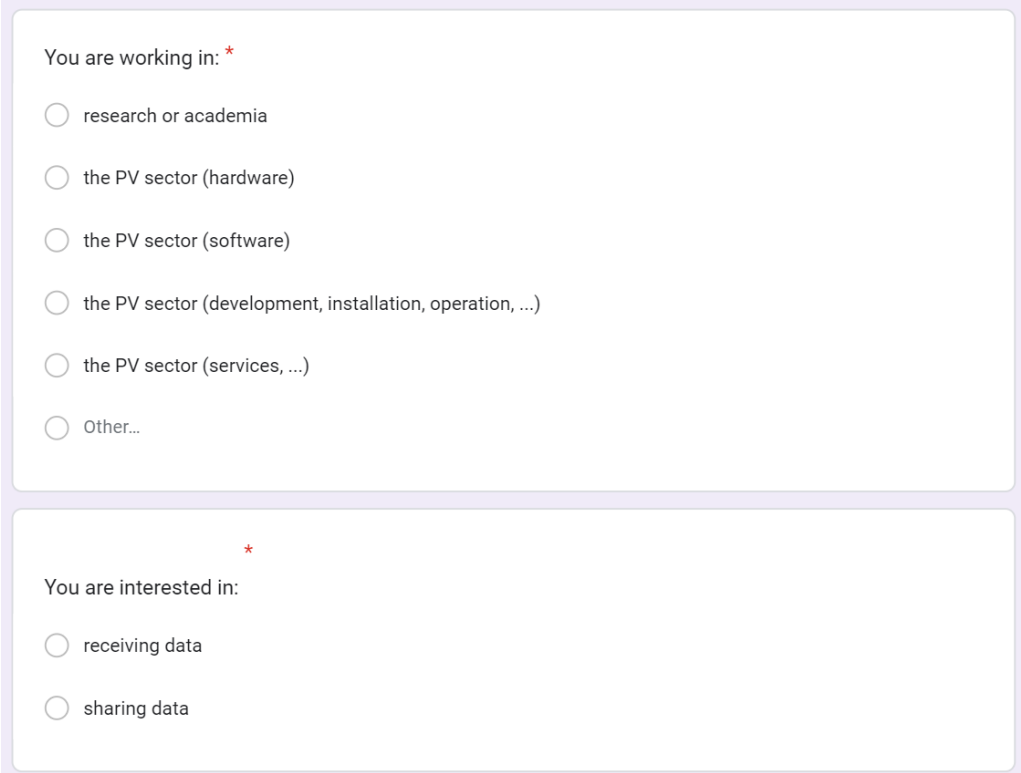
Description (optional)

This data will not be shared and will only be used for the purpose of this work and your answers will be anonymised. You can withdraw your consent to share this data with us at anytime.

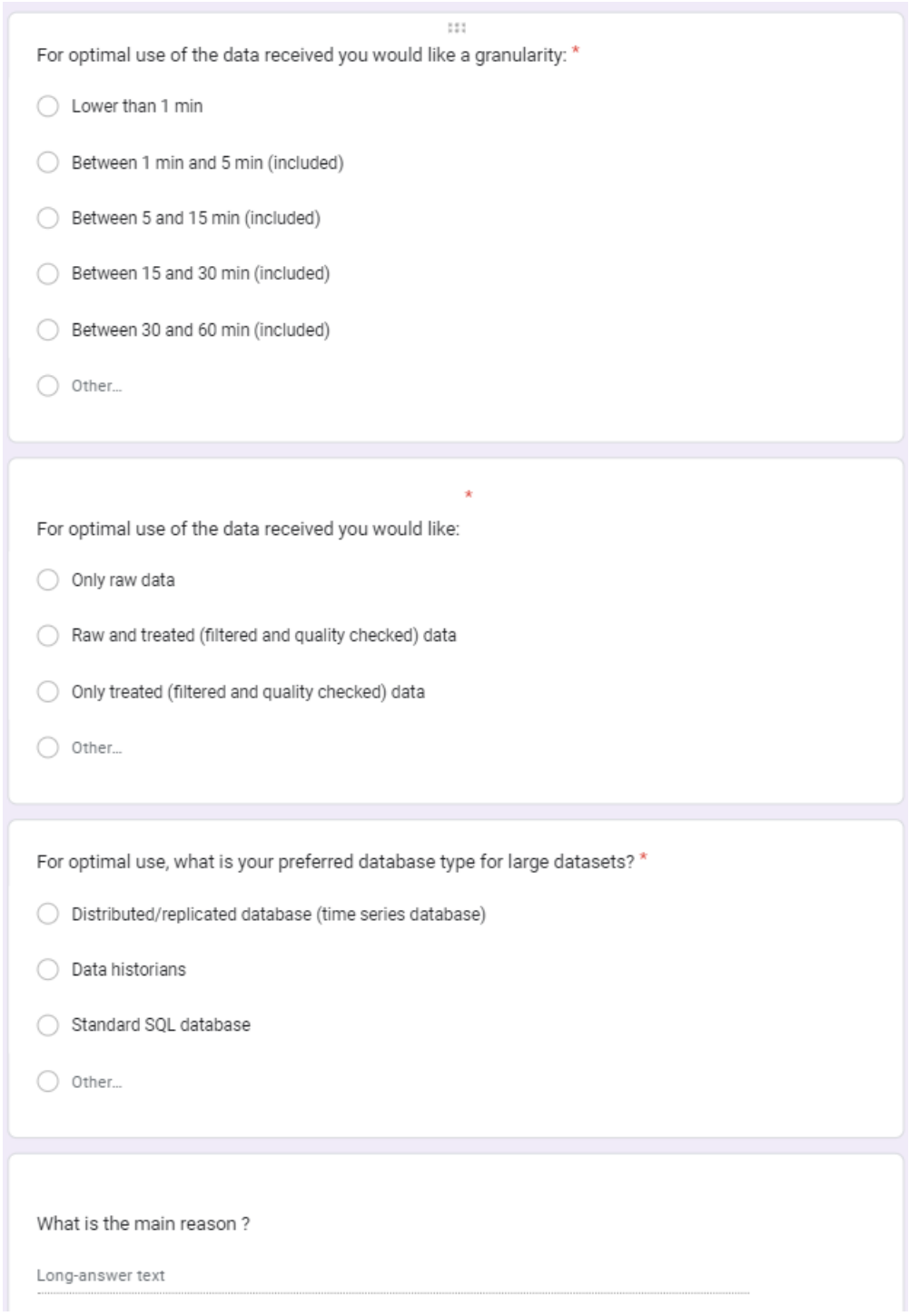
General Data Protection Regulation (GDPR)
SERENDI PV consortium complies with all applicable data privacy laws and regulations including, but not limited to, the General Data Protection Regulation (GDPR). Under the GDPR rules and regulations, you may have certain data rights. If you desire to exercise any of these rights, please send an email to with "Data Privacy Request" in the subject line, and in the body of the email, please specify the precise privacy right for which you need help. Please note that further information may be required before a request can be fulfilled, and SERENDI PV consortium retains the right, where authorised, to impose a fee to cover the expense of some requests.
Contact Us
If you have any questions or concerns about this Privacy Policy or about the use of your personal information, please feel free to contact us by email at contact@coplasimon.eu

Short-answer text

- **Information about surveyed COPLASIMON user**



- **Survey Questions (COPLASIMON user as data receiver)**



For optimal use, what is your preferred database/file format for small datasets ? *

- CSV
- Standard SQL database
- Better file formats – human readable
- Better file formats – non-human readable
- Other...

What is the main reason ?

Long-answer text

For optimal use, what is your preferred data exchange/transfer method? *

- SQL layer for database
- Message queuing system (e.g. pub/sub)
- SFTP
- API (HTTPS REST)
- CKAN (https://ckan.coplasimon.eu/)
- Other...

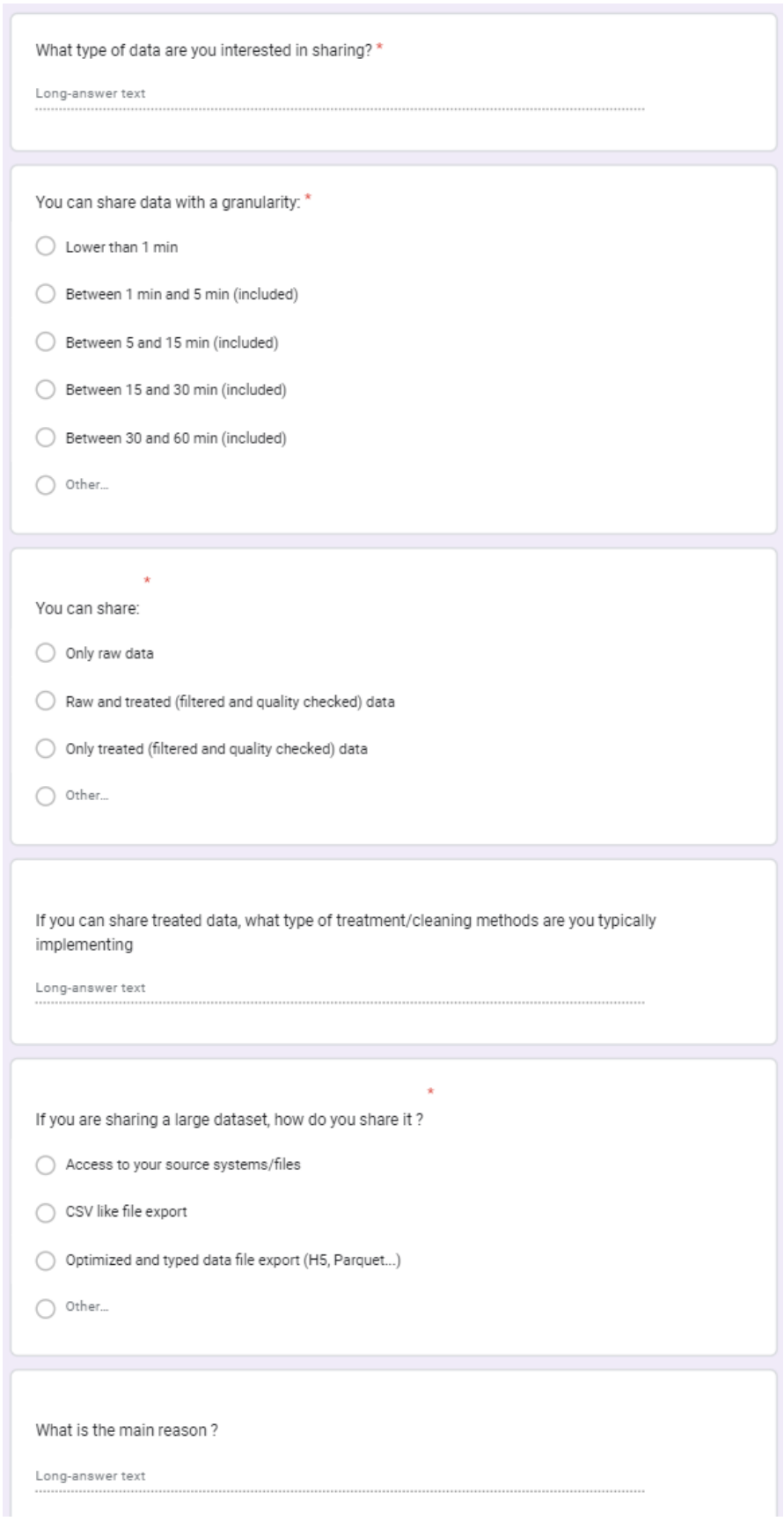
What is the main reason ?

Long-answer text

For optimal use, what is your preferred data privacy method *

- Pseudonymization
- K-anonymity / differential privacy
- Time-series anonymization
- Not applicable
- Other...

- **Survey Questions (COPLASIMON user as data sharer)**



If you are sharing a small dataset, which database/file format is it ? *

- CSV
- Standard SQL database
- Better file formats – human readable
- Better file formats – non-human readable
- Other...

What is the main reason ?

Long-answer text

Which method would you prefer for data exchange/transfer ? *

- SQL layer for database
- Message queuing system (e.g. pub/sub)
- SFTP
- API (HTTPS REST)
- CKAN (https://ckan.coplasimon.eu/)
- Other...

What is the main reason ?

Long-answer text

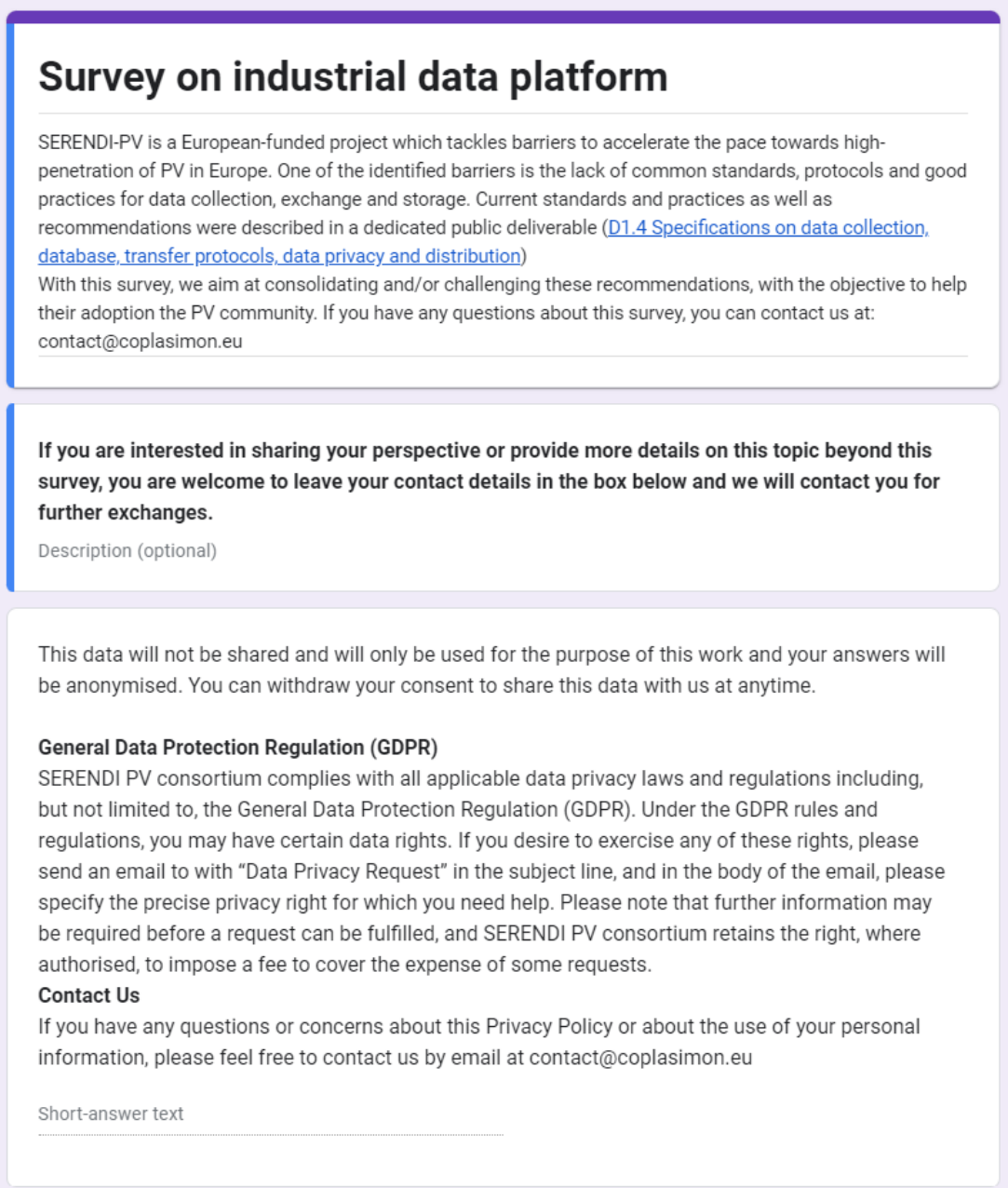
Which data privacy method are you using for the data you would like to share? *

- Pseudonymization
- K-anonymity / differential privacy
- Time-series anonymization
- Not applicable
- Other...

5.2 Survey on data collection, exchange and storage oriented towards external stakeholders.

The survey is constructed as follow

- **Introduction and GDPR aspects**



Survey on industrial data platform

SERENDI-PV is a European-funded project which tackles barriers to accelerate the pace towards high-penetration of PV in Europe. One of the identified barriers is the lack of common standards, protocols and good practices for data collection, exchange and storage. Current standards and practices as well as recommendations were described in a dedicated public deliverable (D1.4 Specifications on data collection, database, transfer protocols, data privacy and distribution)
With this survey, we aim at consolidating and/or challenging these recommendations, with the objective to help their adoption the PV community. If you have any questions about this survey, you can contact us at: contact@coplasimon.eu

If you are interested in sharing your perspective or provide more details on this topic beyond this survey, you are welcome to leave your contact details in the box below and we will contact you for further exchanges.

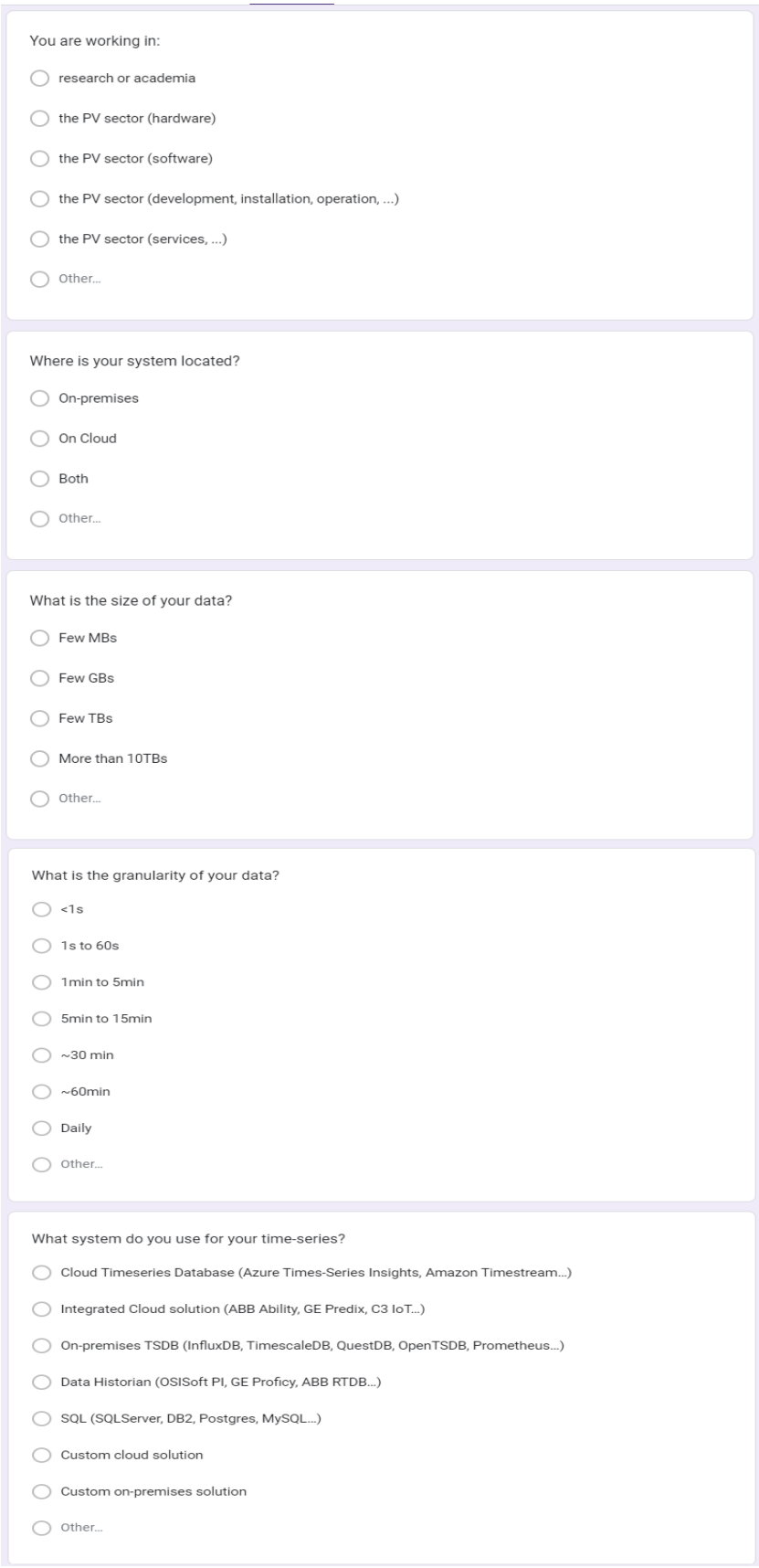
Description (optional)

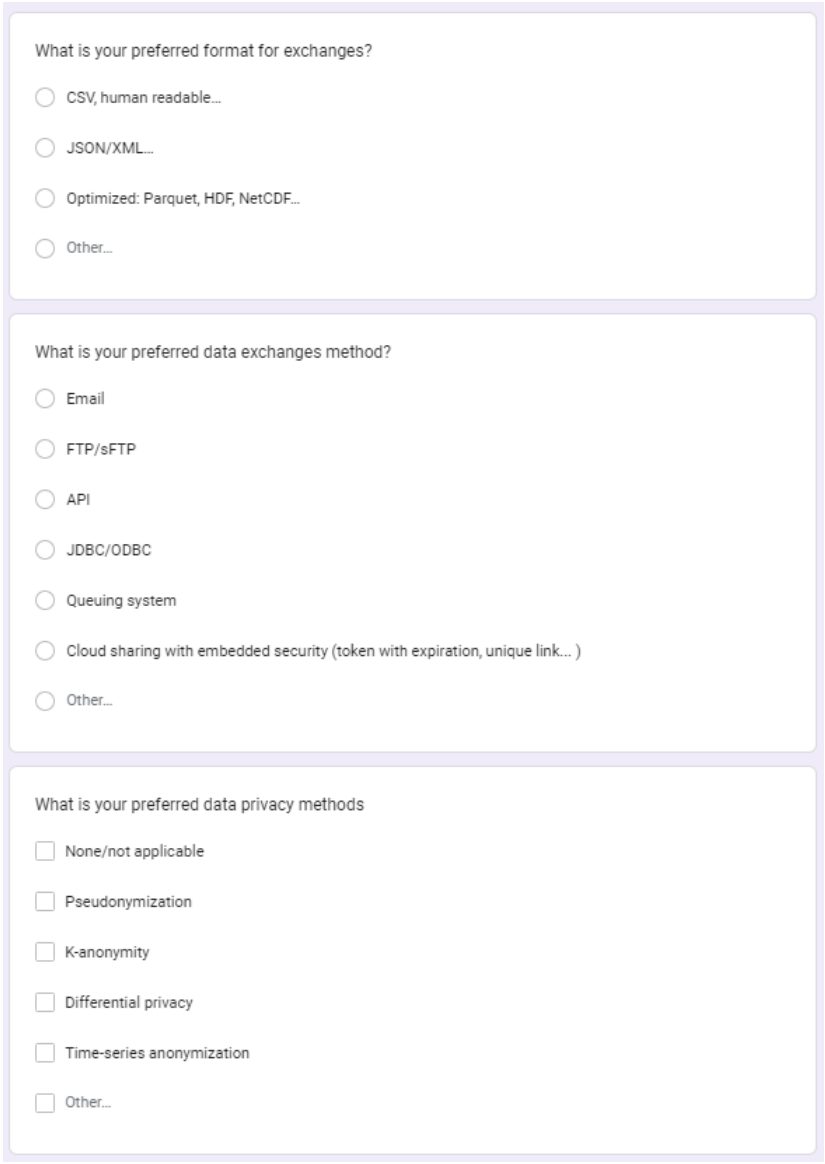
This data will not be shared and will only be used for the purpose of this work and your answers will be anonymised. You can withdraw your consent to share this data with us at anytime.

General Data Protection Regulation (GDPR)
SERENDI PV consortium complies with all applicable data privacy laws and regulations including, but not limited to, the General Data Protection Regulation (GDPR). Under the GDPR rules and regulations, you may have certain data rights. If you desire to exercise any of these rights, please send an email to with "Data Privacy Request" in the subject line, and in the body of the email, please specify the precise privacy right for which you need help. Please note that further information may be required before a request can be fulfilled, and SERENDI PV consortium retains the right, where authorised, to impose a fee to cover the expense of some requests.
Contact Us
If you have any questions or concerns about this Privacy Policy or about the use of your personal information, please feel free to contact us by email at contact@coplasimon.eu

Short-answer text

- **Survey Questions**



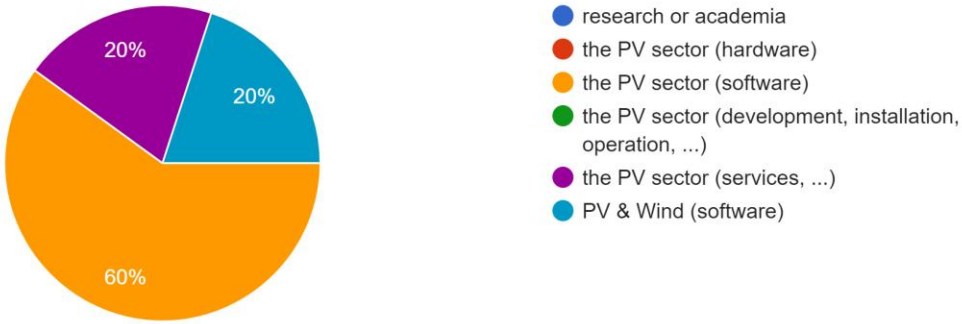

What is your preferred format for exchanges?
CSV, human readable...
JSON/XML...
Optimized: Parquet, HDF, NetCDF...
Other...
What is your preferred data exchanges method?
Email
FTP/sFTP
API
JDBC/ODBC
Queuing system
Cloud sharing with embedded security (token with expiration, unique link...)
Other...
What is your preferred data privacy methods
None/not applicable
Pseudonymization
K-anonymity
Differential privacy
Time-series anonymization
Other...

5.3 Results of the survey on data collection, exchange and storage.

The answers collected as of 25/09/2024 are presented below.

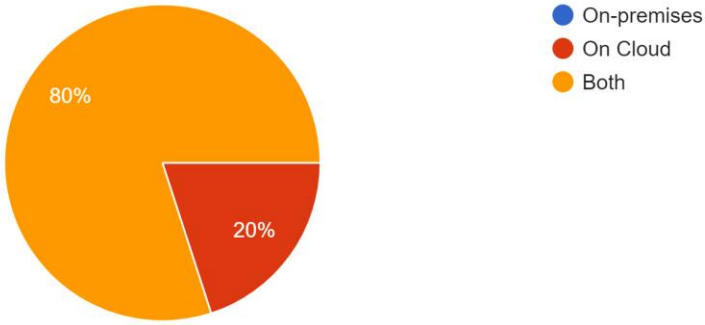
You are working in:

5 responses



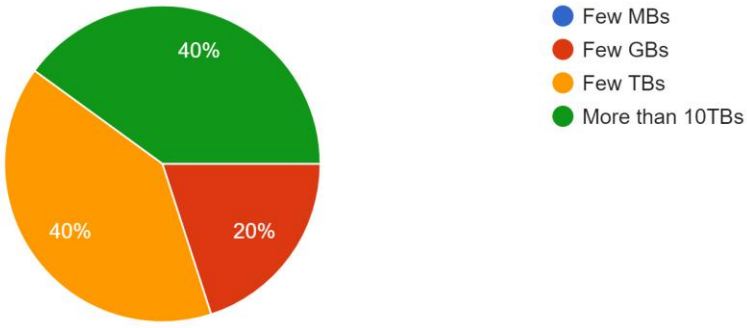
Where is your system located?

5 responses



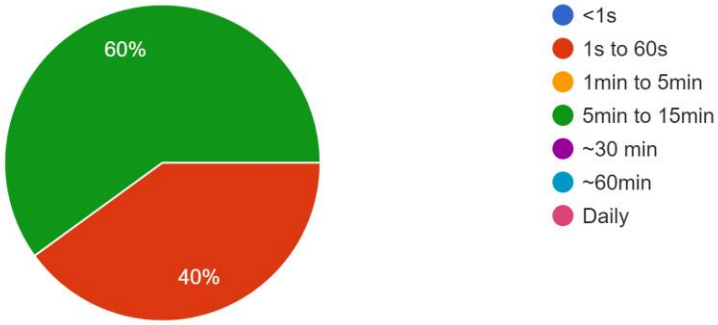
What is the size of your data?

5 responses



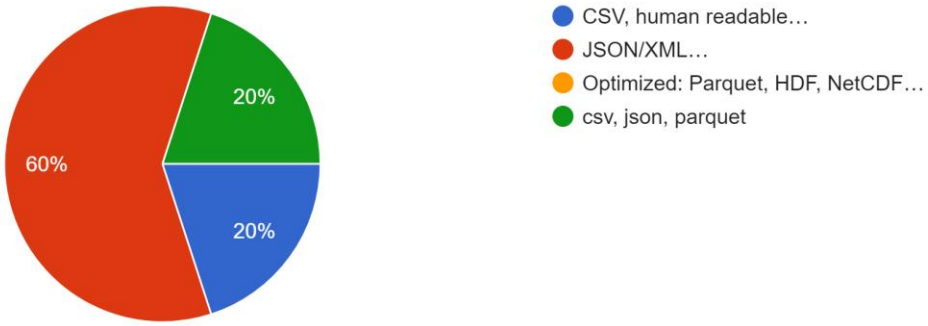
What is the granularity of your data?

5 responses



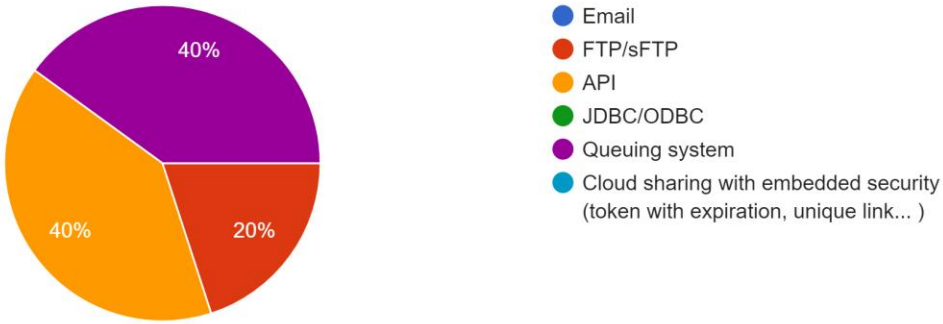
What is your preferred format for exchanges?

5 responses



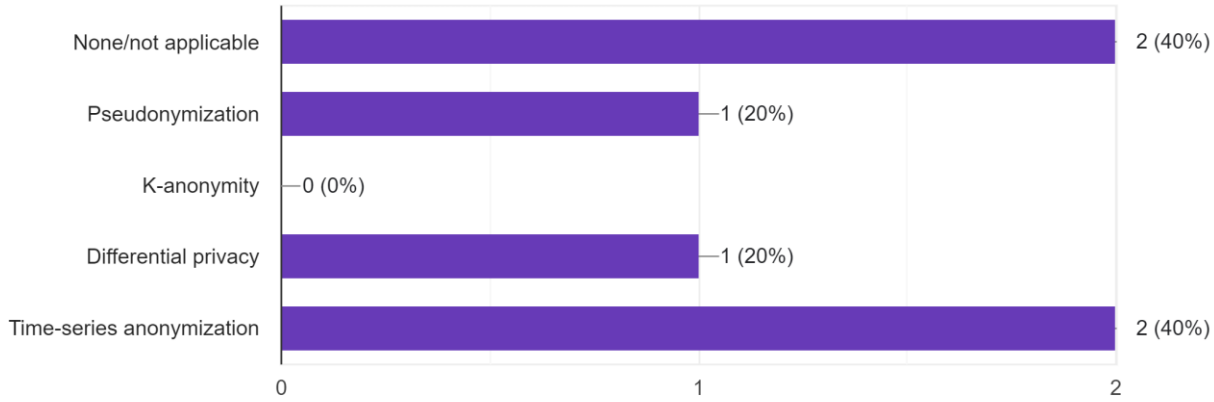
What is your preferred data exchanges method?

5 responses



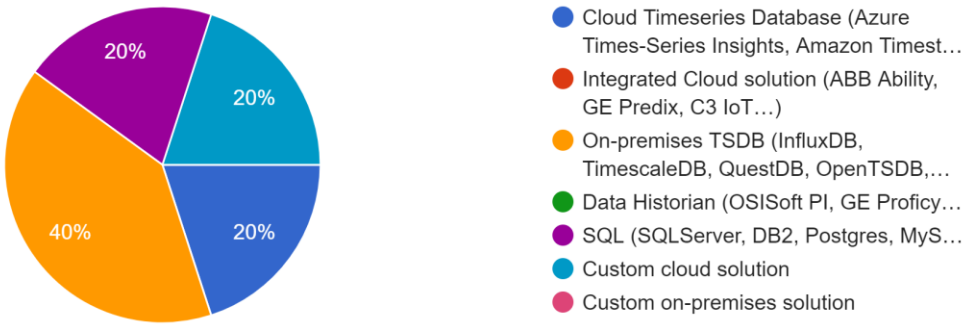
What is your preferred data privacy methods

5 responses



What system do you use for your time-series?

5 responses



6 REFERENCES

[1] Pierre-Jean Alet *et al.*, "Mapping the relevance of digitalization for photovoltaics - ETIP-PV working group Digital PV and Grid," 2023, doi: 10.13140/RG.2.2.36562.12489.

[2] J. Leloux, "Mapping the Relevance of Digitalization for Photovoltaics," 2024, doi: 10.13140/RG.2.2.36739.92962/1.

[3] "Reliability and Performance of Photovoltaic Systems," IEA-PVPS. Accessed: Oct. 01, 2024. [Online]. Available: https://iea-pvps.org/research-tasks/performance-operation-and-reliability-of-photovoltaic-systems/

[4] C. Deline *et al.*, "PV Fleet Performance Data Initiative Program and Methodology," in *2020 47th IEEE Photovoltaic Specialists Conference (PVSC)*, Calgary, AB, Canada: IEEE, Jun. 2020, pp. 1363–1367. doi: 10.1109/PVSC45281.2020.9300583.

[5] "Solar Data Bounty Prize | American-Made Challenges." Accessed: Oct. 01, 2024. [Online]. Available: https://americanmadechallenges.org/challenges/solar-data-bounty

[6] S. Vitale *et al.*, "Shaping European Collaboration on Photovoltaics: A Collaborative Platform For Simulation And Monitoring (COPLASIMON)," Sep. 2024.