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

SERENDI PV

D1.4 Specifications on data collection, database, transfer protocols, data privacy and distribution and Intellectual Property.

T1.4 PV in the digital era: specifications on data collection, database, transfer protocols, data privacy, distribution and IP.

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:	Innovation and Networks Executive Agency (INEA)
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.:	SERENDI-PV_D1.4 Specifications on data collection, database, transfer protocols, data privacy, distribution and IP_v1
Lead Beneficiary:	Mylight Systems
Doc. Dissemination Level:	**PU**– Public
Due Date for Deliverable:	31/03/2022 (M18)
Actual Submission date:	13/04/2022 (M19)
Version	1.0

Summary

This deliverable presents the current standards and practices in place within the photovoltaic domain, and then sets out and describes various best practises, recommendations for future protocols and standards to aim for. The standards cover the following themes: data collection, database storage and format, transfer protocols, data privacy, distribution rights and intellectual property.

Up until now, data collecting, sharing, storage and transfer has been handled on a case by case 'ad hoc' basis, each actor having slightly different 'standards' based on their own needs, 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. Whilst it is almost impossible to define an exact set of one-size-fits-all standards that cover all scenarios, a set of best practises and advice can be formulated in order that the PV community adopts and moves towards these "standards".

This deliverable is an output of task [1.4].

Document Information

Title	PV in the digital era: Specifications on data collection, database, transfer protocols, data privacy and distribution and IP
Lead Beneficiary	Mylight Systems
Contributors	TEC, CEA, FHG, BI, QPV, LUC, SGIS, CYT, NKW, ING
Distribution	PU - Public
Report Name	D1.4 Specifications on data collection, database, transfer protocols, data privacy and distribution and IP_v1

Document History

Date	Version	Prepared by	Organisation	Approved by	Notes
02/12/2021	0.1	J. Reed, C. Salperwyck, All participants	All		
31/03/2022	1.0	J.Reed, C.Salperwyck, All participants	MLS and All	TECNALIA	

Acknowledgements

The work described in this publication has received funding from the European Union's Horizon 2020 research and innovation program 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 because 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

To accelerate the pace towards high-penetration of PV in Europe and apart from the purely technical issues covered by SERENDI-PV, additional accompanying measures need to be set in place to lower the existing barriers to PV development, including:

- Remaining uncertainties on PV reliability, performance, and profitability, increasing project financing costs.
- Limited accuracy of the modelling and data analytics, especially for complex typologies. New products and modelling necessities appear every year, such as: thin-film technologies, bifacial PV, BIPV, floating PV, the aging of PV systems, tracking optimization strategies, energy management and storage for grid-connected installations, or the impact of spectral solar radiation on different PV cell technologies.
- Lack of traceability, transparency, and openness around the simulation software. Most of the simulation software used in the PV industry is commercial product, whose source code is not accessible for assessment. The technical knowledge is spread among the scientific community and the industry.
- Lack of reliable and available data on weather conditions, PV components, PV systems performance, and metadata on PV and grids.
- Lack of proper understanding of the underlying concepts from the user of the toolboxes.
- Lack of common standards, protocols and good practices for data collection, exchange and storage.

This report focuses on the last of these existing barriers, 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 can cause problems later, particularly during data exchanges. In this respect this report aims to act as a go-to guide with a set of best practices and recommendations that the PV community should adopt and move towards so that these practices become standards.

The report is divided into several of the main data "themes" and addresses the current and the ideal "standards" or practices in each.

This report is then followed by task 7.2 which is feedback and lessons learnt on the implementation of the recommendations of this report.

1.2 Reference material

The following documents and resources were used as references at various points of this report.

- IEC 61724-1 standard "Photovoltaic system performance –part 1: monitoring"
  - Outlines terminology, equipment, and methods for performance monitoring and analysis of photovoltaic (PV) systems
- European strategy for Data
  - Site with details and resources on future European data strategy
- Design principals for space data
- Towards a European-governed data sharing space. Enabling data exchange and unlocking AI potential", November 2020

- Article on data sharing
- https://ec.europa.eu/energy/sites/default/files/documents/bridge_wg_data_management_eu_reference_architcture_report_ 2020-2021.pdf
  - Article on the Bridge EU data architecture report
- 8601 ISO Standard: Date and time format

1.3 Relation with other activities in the project

Next table 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
3.1	This task is a state of the art in the automated supervision of photovoltaic installations. The task deals with different topics related to the automated supervision and analysis of PV plants: monitoring, data quality, operation analysis, digital twins, fault detection and diagnosis and integration in O&M. This task reviews the state of the art in recent years in this regard, providing references from both the academic world and industry regulation.
7.1	This task will constitute an input for the collaboration platform for simulation and monitoring (COPLASIMON). The collaborative platform will provide several public databases for PV meta data and operation data, and it will seek the involvement of the stakeholders from the solar energy community towards a better standardization of data exchange formats and transfer protocols.
7.2	Task 7.2 is the natural follow on and the implementation of the ideas and plans laid out in task 1.4. The aim of this task is to define as set of long-term data standards for the photovoltaic domain. It is similar to task 1.4 but more focused on the actual data itself, its quality, interoperability and the ease of which different parties can share and exploit the data. The standards to be addresses are data collection (what data), QC and filtering (how data quality is assured, how it is treated, what tools are available), database and format, transfer protocols, standardisation (standard definitions and formats for variable X) and interoperability (the ease of which data can be used by different parties).
7.3	Sound legal framework is important to transparent and secure collection, processing and dissemination of the data and software tools. This task will create an inventory of legal aspects related to data distribution and use of software tools, such as terms of use, confidentiality, and modes of distribution. We will investigate already existing guidelines in the literature as well as the standard procedures amongst partners already working with data as sources for information for the establishment of the final report.
11.2	D11.2: POPD - Requirement No. 2 4.2 The host institution must confirm that it has appointed a Data Protection Officer (DPO) and the contact details of the DPO are made available to all data subjects involved in the research. For host institutions not required to appoint a DPO under the GDPR a detailed data protection policy for the project must be submitted as a deliverable.

	4.13 In case the research involves profiling, the beneficiary must provide explanation how the data subjects will be informed of the existence of the profiling, its possible consequences and how their fundamental rights will be safeguarded.

1.4 Abbreviation list

Table 1.2: Abbreviation List

Abbreviation	Meaning
AI	Artificial Intelligence
AM	Air Mass – measure of the relative optical path of the sunlight in the atmosphere
API	Application Programming Interface
CSV	Comma Separated Values – file format standard
FTP	File Transfer Protocol
GAIA-X	Not Applicable
GDPR	General Data Protection Regulation
GHI	Global Horizontal Irradiance
GPS	Global Positioning System
HDF	Hierarchical Data Format – optimised file format for large datasets
IDS	International Data Spaces
IDSA	International Data Spaces Association
IoT	Internet of Things
JDBC	Java Data Base Connectivity – protocol to connect to a Database
JSON	JavaScript Object Notation – file and exchange format
NetCDF	Version of the Hierarchical Data Format – optimised file format for large datasets
NILM	Non-Intrusive Load Monitoring
NTP	Network Time Protocol
ODBC	Open Database Connectivity - protocol to connect to a Database
OPC	Open Platform Communications – protocol to communicate with industrial devices
PaaS	Platform as a Service
RMS	Root Mean Square
SaaS	Software as a Service
SAX	Symbolic Aggregate approximation: representation of a time-series as list of symbols
SFTP	Secured File Transfer Protocol
SI	Système International des unités (international units system)

SQL	Structured Query Language – query language for databases
SSH	Secure Shell Protocol – used here to secure data exchange base on strong encryption
STC	Standard Test Conditions
TSBS	Time Series Benchmark Suite
TZD	Time Zone Deviation
UTC	Universal Time Coordinated
XML	Extensible Markup Language - file and exchange format

2 INTRODUCTION

Within the photovoltaic domain there are a wide range of different datasets, data exchanges, use cases, actors, rules, standards, and protocols. Well known datasets include the monitoring of residential photovoltaic data energy production with the aim of optimising energy consumption, production (or irradiance), prediction datasets for planning energy availability, or the various information needed from PV plants used to balance the electricity grid. Alongside each of these datasets as well as there being a host of metadata, needed to further understand the main data, there is also a wide range of different transfer protocols, data privacy concerns and the question of who owns what.

As such it is difficult to define a one-size-fits-all set of standards. Currently dataset creation and data exchanges are handled in a sort of 'ad-hoc' case by case basis. This can lead to problems later on for a range of reasons, from mistaken units, misunderstood treatments for data gaps, data privacy issues, data transfer incompatibilities or badly defined distribution rights.

Therefore, in this report, a set of "ideal" protocols and recommendations are defined that should be considered as a basis for all work within the domain, acting as a point of reference for all parties.

Whilst defining a fixed set of standards for all scenarios and actors would be ideal, the following schema shows the complication with doing this. Within the photovoltaic domain, there are a vast array of different actors with different data needs, requirements, use cases, different dataset sizes, different financial capabilities, and different visions and priorities of data and its uses. This is only going to continue with the advancement of technology and the growth of the PV domain and market, and the further integration of PV into the electricity grid on a global scale. As such this report aims to define a set of "ideal" recommendations and options that the community should aim to move towards.

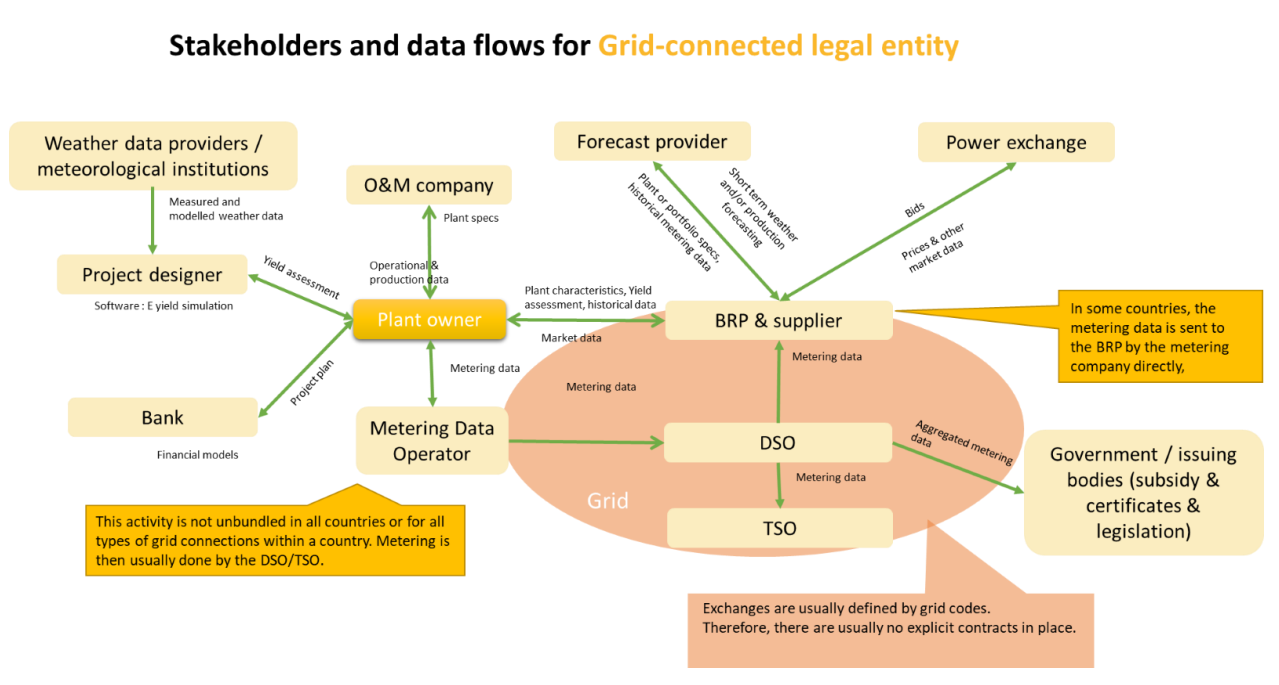


Figure 2.1: Schema PV actors and interactions

Schematic showing some of the different actors and interactions concerning data exchange within the photovoltaic community.

(From Task 7.3)

3 FIELD FEEDBACK AND RECOMMENDATIONS

This section will discuss the various themes of the report, describing current state and standards, various feedbacks and then discuss some recommendations to be implemented throughout the course of the SERENDI-PV project, and to be reviewed in task 7.2. The following themes will be covered: data collection, types of data, data exchange, database, transfer protocols, data privacy, and distribution and IP rules.

Data collection

The collecting, measuring and analysis of data. Under this umbrella we also include quality assurance and control (although this will be addressed in more detail in task 7.2), and formatting. Data collection and what data is collected is the most basic part of all.

As data may come from different monitoring units with independent timestamps, the first key issue is to ensure that all the data are synchronized, preferably by an automated mechanism such as global positioning system (GPS) or network time protocol (NTP).

Data files consist of a first line of headers and a line of measurement values for each time step.

Data file name

The format of the data collected, and its storage will be discussed further in section 3.4 (Database).

As a rule, when sharing a data file, the name should contain a reference to the contents (I.e., “production_monitoring_energy”) and a reference to the date interval (I.e., “23-11-1990__12-12-1994").

Data header and fields definition

Data headers are made up of several fields which should always be in standard English (“active_power_kW”, “energy_kWh”) + separator

The first field is logically “date and time”. It may happen that date and time are separate, but it is advised to use a single representation (see below).

The following fields are dedicated to physical measurements that should be rather explicit and understandable, with a “data name” (with indices if several measurements of the same type are available) and its “unit” written in brackets. The use of special characters such as ^{2}, $^{-1}$ or ° should be avoided as they may be difficult to read.

In csv, xls or txt files, for instance, fields should be separated with the semicolon character (“;”) and not the slash (“/”) one as the latter is sometimes used in units (ex: W/m^2).

Table 3.1: Definition of Fields included in the data headers

Field	Definition	Ideal Frequency and T Horizon	Notes
id	A unique number or string to identify the sensor/site	NA	
Datetime	The time and date of the measurement (format defined below)	In function of use case. 1-minute resolution usually most useful; can then be grouped by larger intervals	

Date	The date of the measurement (without hour/minute)	Daily.	
Energy	Energy value of the measurement (ideal units: kWh)	kWh reading at intervals not equal to an hour must be explicitly stated.	Although not a System International (SI) unit, the kWh is a convenient unit because a power of 1 kW during a time interval of 1-hour results in an energy production of 1 kWh. It is therefore often used in the electric power and photovoltaic energy industries.
Active Power	Effective power consumed by AC circuit (also known as true or real power) (ideal units: kW)		
Reactive Power	Unused power in AC circuit. (ideal units: kVAr)	Depending on use case. 1-minute resolution usually most useful; can then be grouped by larger intervals	
Apparent Power	Product of Root Mean Square (RMS) value of Voltage and Current.	Depending on use case. 1-minute resolution usually most useful; can then be grouped by larger intervals	
Voltage	The difference in electric potential between two points, which is defined as the work needed per unit of charge to move a test charge between the two points. (units: V)	Depending on use case. 1-minute resolution usually most useful; can then be grouped by larger intervals	
Current	The net rate of flow of electric charge through a surface (units: A)	Depending on use case. 1-minute resolution usually most useful; can then be grouped by larger intervals	
Frequency	The frequency of the AC current oscillations (units: Hz)	Depending on use case. 1-minute resolution usually most useful; can then be grouped by larger intervals	
Inclination	Also often called tilt or slope. The angle relative to the horizon [90 degrees is straight up] (units: degrees)	NA	
Orientation	Also often called azimuth. The angle of rotation relative	NA	

	to a reference (0°), which for solar energy applications is often taken as facing equator, with for the northern hemisphere is due south, and due the northern hemisphere is due north (units: degrees)		
Latitude, longitude	Standard coordinates. More decimal points = higher precision [4 or 5 decimal points = accuracy required for solar panels on a residential roof] (units: degrees of latitude and longitude)	NA	High precision coordinates can be considered confidential data as can be used to identify house and, in combination with other data, be used to profile residents.
Installed power (nominal power)	How much power the panel can deliver under Standard Test Conditions (STC). (units kWp).	NA	• solar irradiance of 1,000 W/m² • an ambient temperature of 25°C (an increase in PV cell temperature leads to lower PV output) • Spectral distribution of solar irradiance corresponding to Air Mass 1.5: ASTM G-173-03 (Gueymard, 2004).

Data Format

Format of datetime

Time usually refers to local time but it implies winter/summer time changes, and therefore **universal time is highly recommended**, such as the ISO 8601 with the time zone that should be used when using a "text field" such as YYYY-MM-DDThh:mm:ssTZD (ex: 2020-03-20T01:31:12.467113+00:00), where:

- YYYY = 4 digits for the year
- MM = 2 digits for the month (01=January, etc.)
- DD = 2 digits for the day of the month (01 to 31)
- hh = 2 digits for the hour (00 to 23)
- mm = 2 digits for the minutes (00 to 59)
- ss = 2 digits for the seconds (00 to 59)
- s = one (or more) digit(s) representing a decimal fraction of the second
- TZD = Time Zone Deviation (Z or +hh:mm or -hh:mm, thus the deviation from GMT time).

If the date is represented as number, it should be with the "classical" number of milliseconds between the time to store and midnight, January 1, 1970 UTC (universal time coordinated). The ISO 8601 (Data elements and interchange formats – Information interchange – Representation of dates and times) format with the time zone is especially easy to parse in today's programming languages used in the community (e.g., Python pandas).

There is also ISO 8601 standard for periods of time – durations (https://en.wikipedia.org/wiki/ISO_8601#Durations). This can be used in data requests when requesting certain data granularity.

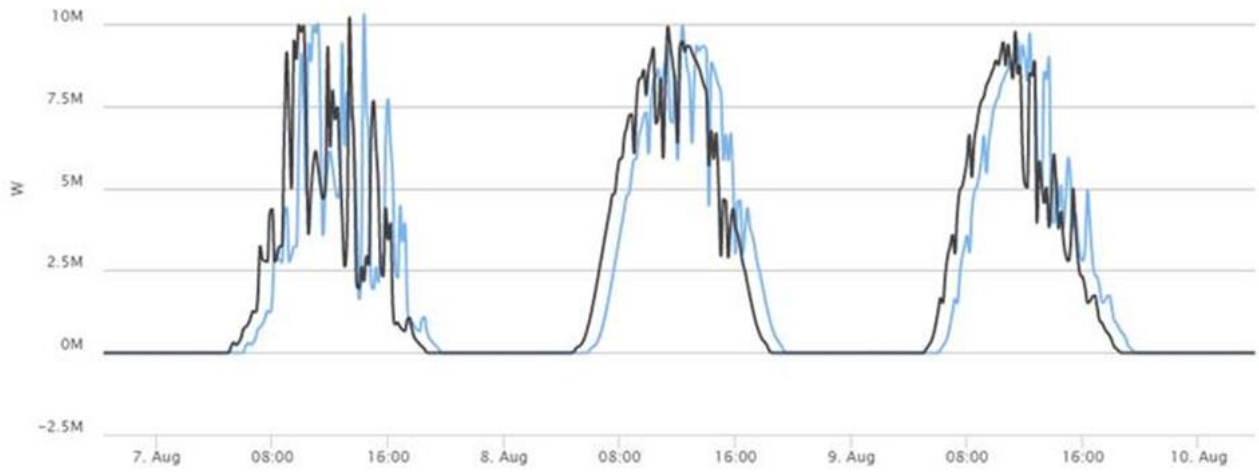


Figure 3.1: Local Time – Universal Time deviation example

Example PV plant active power measured by 2 monitoring systems, one using local time and the other using universal time, leading to 1h deviation 6 months a year

Format and unit of numerical data

The decimal character **should be a dot** and not the comma. The unit should be clearly indicated in the header.

It is also important to know if the value given is for the given timestamp or for the time interval between two timestamps. For example, at Solargis, for sub-hourly data the simulated PV output value typically represents instantaneous value (a power measured exactly at the given timestamp), so the PV output unit will be the kW (or W/m^2 in case of solar irradiation). For hourly and bigger aggregation, the value represents an energy accumulated or averaged within the aggregated interval - the unit is the kWh (or Wh/m^2, kWh/m^2 in case of solar irradiation).

Data precision

Numerical values should be written with at least three digits after the dot. In the following table, all columns conform to this standard except GHI1 (Global Horizontal Irradiance) which is limited to two decimal points.

Date_time	Tair (C)	GHI1 (W.m-2)	GPOA1 (W.m-2)	GPOA2 (W.m-2)	Inv1_DC_A1 (A)	Inv1_DC_A2 (A)	Inv1_DC_V1 (V)
2016-09-08T10:10:10Z	20.800	756.56	312.458	310.686	2.943	4.138	232.990
2016-09-08T10:10:15Z	20.800	755.94	312.421	310.662	2.933	4.127	233.012
2016-09-08T10:10:20Z	20.800	755.97	312.201	310.415	2.923	4.123	232.872
2016-09-08T10:10:25Z	20.800	755.76	311.742	309.926	2.914	4.112	232.952
2016-09-08T10:10:30Z	20.900	756.54	311.121	309.296	2.914	4.091	232.998
2016-09-08T10:10:35Z	20.850	754.43	310.280	308.420	2.900	4.070	233.074
2016-09-08T10:10:40Z	20.850	755.81	309.195	307.369	2.885	4.052	232.806
2016-09-08T10:10:45Z	20.850	755.1	307.885	306.076	2.867	4.034	233.115
2016-09-08T10:10:50Z	20.850	755.46	306.517	304.739	2.867	4.023	233.212
2016-09-08T10:10:55Z	20.850	753.34	305.055	303.291	2.846	3.999	233.240
2016-09-08T10:11:00Z	20.850	753.38	303.628	301.839	2.829	3.974	233.154
2016-09-08T10:11:05Z	20.850	754.89	302.196	300.400	2.810	3.952	233.164
2016-09-08T10:11:10Z	20.850	754.41	300.578	298.870	2.789	3.924	233.136
2016-09-08T10:11:15Z	20.850	753.36	298.921	297.127	2.789	3.909	233.066
2016-09-08T10:11:20Z	21.000	752.51	296.980	295.258	2.767	3.884	233.060
2016-09-08T10:11:25Z	21.000	752.37	294.976	293.282	2.747	3.865	233.234
2016-09-08T10:11:30Z	21.000	753.98	293.187	291.480	2.728	3.851	233.317

From the data storage point of view, all data have to ensure high accuracy. Decimal numbers should be stored as float (32-bit IEEE 754). Integer numbers should ensure the full range of the measure. Another possibility is to choose a unit that would not need to use decimal values. The advantage of using integers is that there is no side effect of floats that sometimes represent an integer as float with lots of “9” or “0” (4 being represented as 3.999999 for example). For accounting/billing purposes, floats have the same drawbacks: sum of many floats are sometimes different from the sum of many integers.

From the visualization point of view, there are numeric rules about this topic. The most common limit the number of ciphers:

- A decimal number will have a maximum of X ciphers. The decimal part will be cut to accomplish this rule.
- The integer part will be represented completely.
  For example, if the number is limited to 4 ciphers, and the number is 123,456 it will be represented as 123,4. If the number is 123456,7 it will be represented as 123456.

Data granularity

It really depends on the use case. Perhaps the most practical approach is to not expect (or store it in a metadata) the harmonic time step (where data is regularly spaced with a fixed time-step and no missing data) when ingesting raw input data from sensors. Taking the data points as a stream of events with just removing duplicates (which means updating the existing values). The prevailing time step can be detected from timestamp deltas. In later processing steps the gaps in time (i.e., completely missing timestamps) are detected by comparing raw data towards the desired harmonic time step.

Electrical data, such as current, voltage and power at inverter level or at the point of delivery are often available at a 10- or 15-minute recording time step, based on the average of sub-minute or second samples. The recording interval is thus often the latter for the whole PV plant.

But it can be useful to get data at a thinner granularity, with a 1-minute recording interval, for irradiation measurement for instance to better take into account the irradiance variability. In this case, some lines of the data file are to be empty regarding the inverter for instance.

The recorded value can be the average, minimum, maximum, cumulate or other of the samples over the recording interval, depending on the measurement type.

In general, a granularity between 1 and 5 minutes (favouring 1 minute) is sufficient for most analyses. The data can then be aggregated over larger periods for other analyses, storage, or transfer as required.

Data quality

We should first distinguish the raw data file (with no treatment), provided by the acquisition system, and the modified data file, made after filtering, abnormal value detection and correction. Any data treatment should be noted and shared with the sharing of data to avoid problems and confusion later. A good standard practice is to keep a raw and treated version of all data.

Or, for any dataset, via an additional column, a column value can have an associated 'flag' code describing the value origin, e.g., original value, missing original value filled by the calculated value, the modelled value from different source or long-term average values and so on. These origin flag values can be mixed with 'standard 'flag values coming from quality control procedure, e.g., valid values, anomalous values, etc. Although the separation of quality flag from the origin flag would be clearer.

Anomalous values?

The best option for anomalous values is to remove them and replace by a Nan. Any treatment (I.e., smoothing) introduces a bias. Whatever approach is chosen, the methodology should be noted and shared with any sharing of the data.

How are data gaps treated?

Data gaps should be either not included (I.e., no data 'row') or filled with NaN. This signifies an absence of data rather than the ambiguity of 0.

However, data gaps can also be filled in via imputation. The missing data may be (re)calculated by estimates based on:

- The previous or following (good) values when it is a one-time error,
- An average value over a short period of time
- Values in similar conditions of operation (of the same measurement itself or measurements on other components)

But the uncertainty increases with the duration of missing data, and the best solution is often to discard these data from the analysis. More detail on the treatment of anomalous values and data gaps can be found in section 2 of the *D3.1: Revision of state of the art of automatic PV performance supervision systems*.

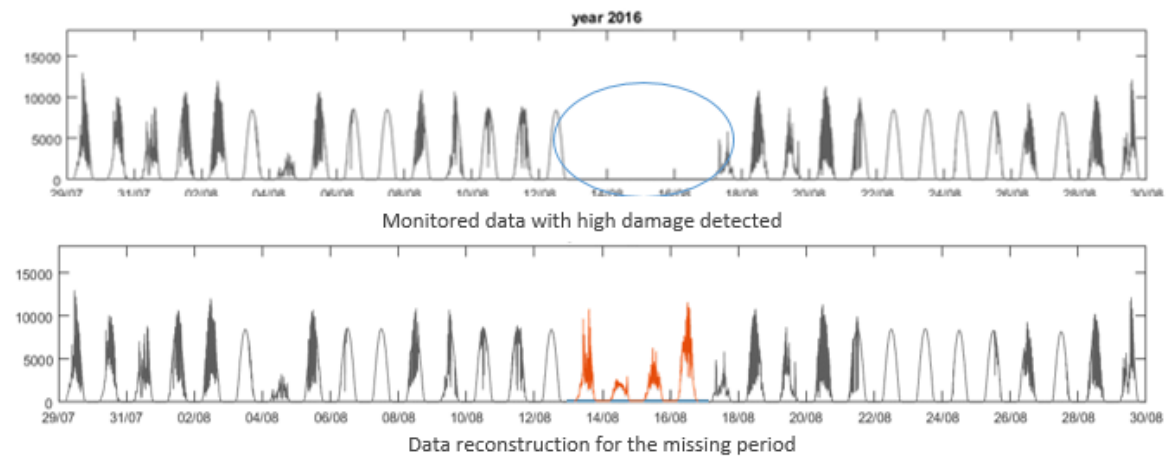


Figure 3.2: Data Reconstruction example

Example of data reconstruction for a missing time period based on values in similar operating conditions.

Data Analysis

There are a wide range of standard techniques for the use and analysis of photovoltaic data. Large amounts of information and detail can be found in the *D3.1*.

3.1 Data

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 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. Hourly 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 a time and production may be required.

3.1.1 Production Data

It can likely be agreed that one of the fundamental datasets in the photovoltaic domain is that of the production data of the photovoltaic installation itself. However, this dataset can be interpreted and presented in many ways and perfectly illustrates the need for standardisation.

	date	datetime	user_id	prod	load	grid	injection	devices
1	2021-09-01	2021-09-01T04:30:00.000+0000	0NeLI0kUmkD4PB0C	0	0.051612124	0.051612124	0	{"water_heater": {"F8DFE20103F1-2": 0}}
2	2021-09-01	2021-09-01T07:00:00.000+0000	0NeLI0kUmkD4PB0C	0.07292233	0.054147992	0	0.01877433	{"water_heater": {"F8DFE20103F1-2": 0}}
3	2021-09-01	2021-09-01T00:45:00.000+0000	0NeLI0kUmkD4PB0C	0	0.35173368	0.35173368	0	{"water_heater": {"F8DFE20103F1-2": 0}}
4	2021-09-01	2021-09-01T10:15:00.000+0000	0NeLI0kUmkD4PB0C	0.5187157	0.38625687	0	0.13245882	{"water_heater": {"F8DFE20103F1-2": 0}}
5	2021-09-01	2021-09-01T07:45:00.000+0000	0NeLI0kUmkD4PB0C	0.1521044	0.04711267	0	0.104991734	{"water_heater": {"F8DFE20103F1-2": 0}}
6	2021-09-01	2021-09-01T00:00:00.000+0000	0NeLI0kUmkD4PB0C	0	0.0460593	0.0460593	0	{"water_heater": {"F8DFE20103F1-2": 0}}
7	2021-09-01	2021-09-01T03:15:00.000+0000	0NeLI0kUmkD4PB0C	0.0029752862	0.78242576	0.7794505	0	{"water_heater": {"F8DFE20103F1-2": 0.7359072}}
8	2021-09-01	2021-09-01T14:45:00.000+0000	0NeLI0kUmkD4PB0C	0.36773843	0.056338176	0	0.31140023	{"water_heater": {"F8DFE20103F1-2": 0}}
9	2021-09-01	2021-09-01T22:45:00.000+0000	0NeLI0kUmkD4PB0C	0	0.04437272	0.04437272	0	{"water_heater": {"F8DFE20103F1-2": 0}}
10	2021-09-01	2021-09-01T11:00:00.000+0000	0xxjh0GYHKAJqN5t	0.57629937	0.93255687	0.35625747	0	{"water_heater": {"F8DFE2011E44-2": 0.77701926}}
11	2021-09-01	2021-09-01T02:45:00.000+0000	0xxjh0GYHKAJqN5t	0	0.708503	0.708503	0	{"water_heater": {"F8DFE2011E44-2": 0.68776125}}
12	2021-09-01	2021-09-01T18:15:00.000+0000	0xxjh0GYHKAJqN5t	0.0000057666666	0.13232175	0.13231598	0	{"water_heater": {"F8DFE2011E44-2": 0}}
13	2021-09-01	2021-09-01T22:15:00.000+0000	0xxjh0GYHKAJqN5t	0	0.06304416	0.06304416	0	{"water_heater": {"F8DFE2011E44-2": 0}}
14	2021-09-01	2021-09-01T21:30:00.000+0000	0xxjh0GYHKAJqN5t	0	0.19108433	0.19108433	0	{"water_heater": {"F8DFE2011E44-2": 0}}
15	2021-09-01	2021-09-01T20:45:00.000+0000	0xxjh0GYHKAJqN5t	0	0.048232906	0.048232906	0	{"water_heater": {"F8DFE2011E44-2": 0}}
16	2021-09-01	2021-09-01T21:15:00.000+0000	0xxjh0GYHKAJqN5t	0	0.07299279	0.07299279	0	{"water_heater": {"F8DFE2011E44-2": 0}}
17	2021-09-01	2021-09-01T21:45:00.000+0000	0xxjh0GYHKAJqN5t	0	0.03807992	0.03807992	0	{"water_heater": {"F8DFE2011E44-2": 0}}
18	2021-09-01	2021-09-01T06:45:00.000+0000	1AqRKPHcFelJabYn	0.15832736	0.09896897	0	0.059358392	{"water_heater": {"F8DFE2011432-2": 0}}
19	2021-09-01	2021-09-01T05:15:00.000+0000	1AqRKPHcFelJabYn	0.031436663	0.051298928	0.028182596	0.008320331	{"water_heater": {"F8DFE2011432-2": 0}}

Figure 3.3: PV Data example

Example of Photovoltaic installation monitoring data format from Mylight Systems aggregated at 15-minute intervals.

Contains:

- date (datetype)
- datetime (timestamptype)
- user_id (string) [installation identification]
- prod (float) [kWh] The production of the installation
- load (float] [kWh] The consumption of the installation
- grid (float) [kWh] What is being taken from the electricity grid (load – prod)
- injection (float) [kWh] What is being injected into the electricity grid (prod – load)
- devices (json) [kWh] If data for specific appliances required, stored in json

In a separate table using the user_id as a foreign key, all metadata for each installation is available.

3.1.2 Grid Exchange Data

Grid exchange data is the data used to measure what is being drawn from and injected into the electricity grid. This data is important and can have ramifications both at a local, regional, national, and even international scale (for example the European electric grid). A smart metering system is an electronic system capable of measuring electricity fed into the grid, or electricity consumed from the grid, providing more information than conventional meters. Such system is capable of transmitting and receiving data for information, monitoring and control purpose, using a form of electronic communication and comes with a range of benefits for the energy system and its users. Grid data is provided mainly by system operators and utilized by either system operators and aggregators, consumers or producers. System operators exchange grid data for coordinating system operation and planning. The other stakeholders require grid data to use their flexibility in a grid-friendly manner. For example, if there is congestion within the distribution system, an aggregator or distributed generator could be informed about this congestion and adjust its production or consumption accordingly (Thema Consulting Group, 2017).

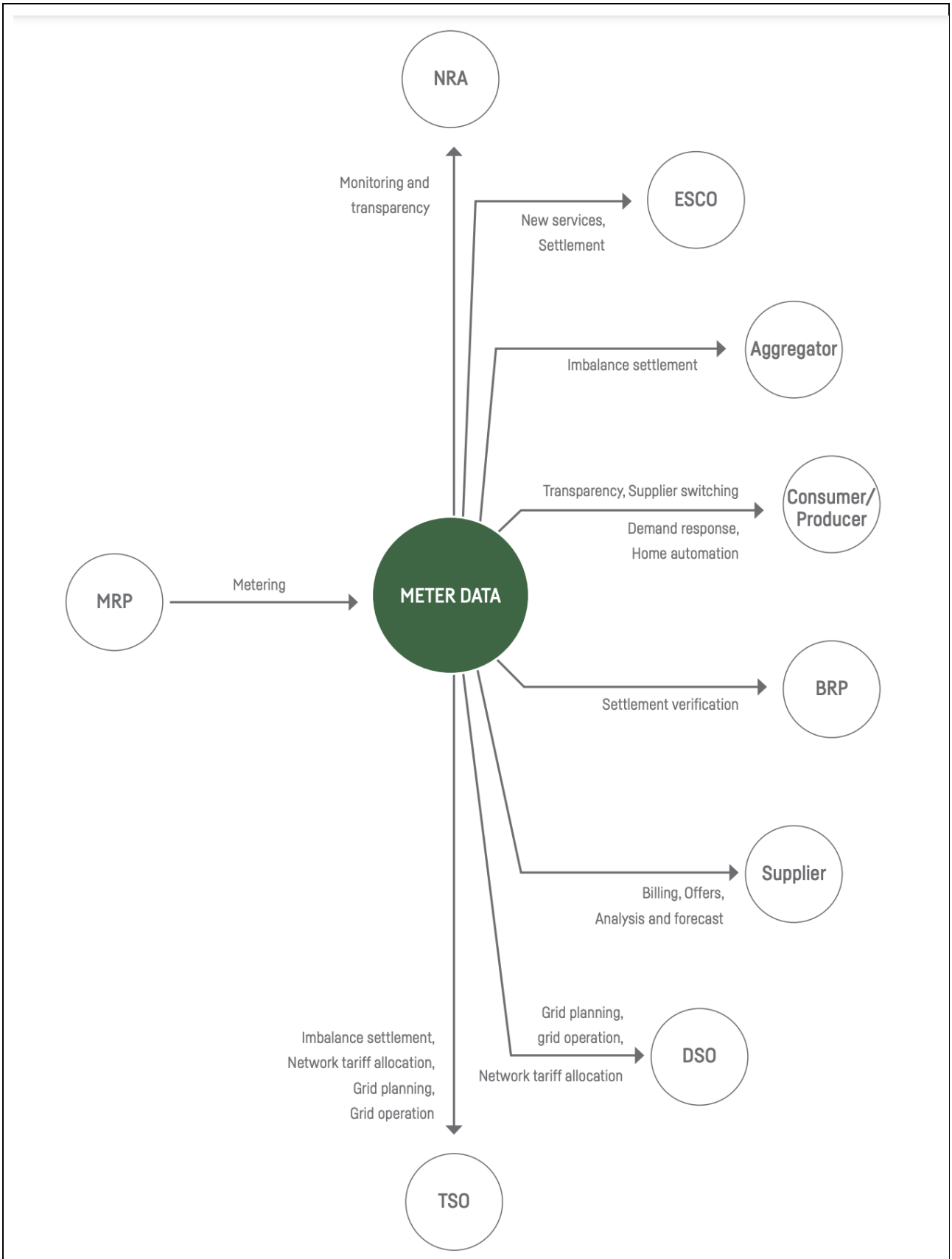


Figure 3.4: Metering Data Needs

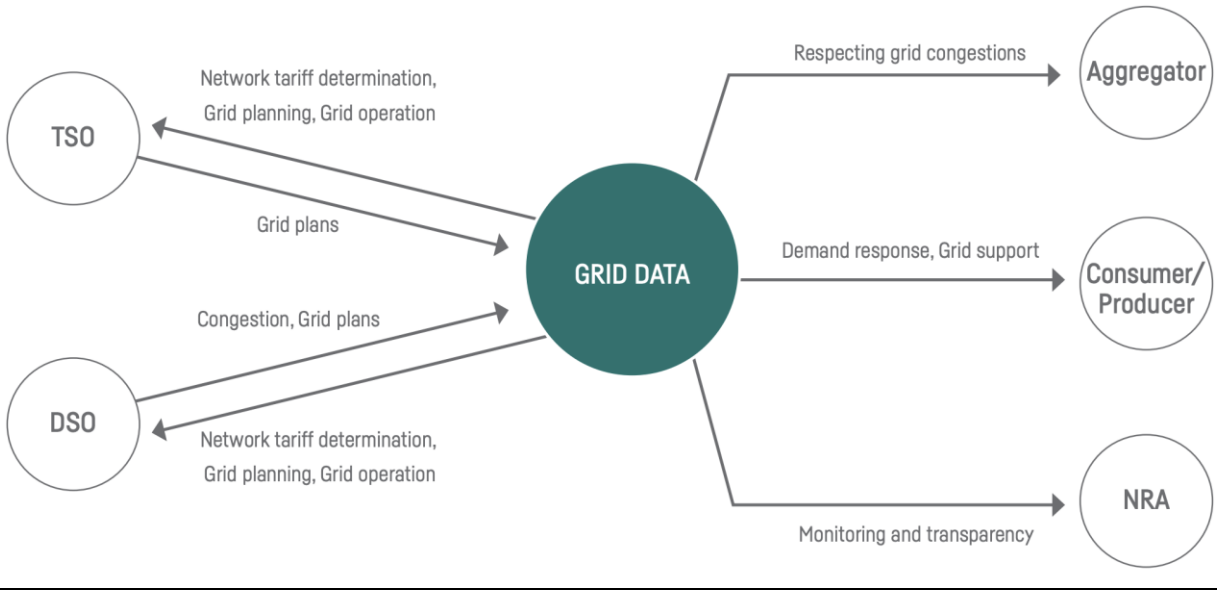


Figure 3.5: Grid Data Needs

In Slovakia, data received are readings of the electricity meter in kW with frequency of 15 minutes from customers (automated email delivery). Customers also send exclusion time slots, when the PV power plant was not in operation due to switch-off by the distribution company. Available are also monthly summaries of invoiced power production, which should match with the 15-minute readings. All performance evaluation is based on these inputs (plus related metadata) combined with simulated data.

In France, consumption and grid exchange data is measured by a monitoring device known as a 'linky'. Electricity consumption and production data is monitored at daily and 30-minute intervals. It is possible to measure both what is being withdraw from the grid, and what is being injected.

3.2 Data exchange

As explained in previous sections, data availability and reusability are one of the main blockers for digitalisation in the PV industry. It is believed that technological solutions that allow the sharing of data and data analytics services between companies could help to overcome the current barriers. In fact, the development of data sharing solutions opens up a new range of possibilities: on the one hand, it allows access to a wider range of operating data with which to improve current models. In addition, it allows the development of solutions to problems in which many actors such as the electricity system intervene. On the other hand, it allows companies to access a global market of providers and consumers of data and services (Accenture, 2018; Otto, Lis, & Cirullies, 2019)

In this sense, the European Commission has published the "European Data Strategy" with which it seeks to make the European Union a leader in a data-driven society and in which it intends to create a single data market in which data can circulate throughout the union and between sectors, for the benefit of all. In this European data market, all European standards, in particular on privacy and data protection, as well as competition law, must be fully respected. It also establishes that the rules for access to data and its use are fair, practical, and clear.

The most pressing inter-organisational concern remains the lack of functional and trustworthy data sharing ecosystems that inspire immediate large-scale participation. Primary causes include the 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 compounded by the lack of widespread adherence to emerging best practices and standards (e.g., interoperability, provenance, and quality assurance standards), whose maturity pace also continues to fail expectations. From a technical point of view, data sharing solutions need to better address European concerns like ethics-by-design for democratic AI, and the rapid shift towards decentralized mixed-mode data sharing and processing architectures also poses significant scalability challenges.

In terms of intra-organisational concerns, a first major concern is the difficulty to determine the value of data, due to a lack of data valuation standards and assessment tools, compounded by the highly subjective and party-dependent nature of data value and the lack of data sharing foresight exhibited by a majority of producers. The second concern revolves around the difficulty faced by data producers balancing their data's perceived value (after sharing) against risks exposed (upon its sharing) despite adhering to standard guidelines. Specific examples include the perceived loss of control over data (due to the fluid nature of data ownership, which remains hard if not impossible to legally define), the loss of trade secrets due to unintentional exposure or malicious reverse-engineering (in a business landscape that is already very competitive), and the risk of navigating around legal constraint in view of potential data policies breaches (including GDPR and exposure of private identities).

When developing platforms that allow the sharing of data and services, there are two main approaches: 1) centralized platforms and 2) decentralized (federated) platforms.

Centralized platforms are solutions in which all data is centralized in a single repository or Data Lake and all analytics services are deployed on the same platform. Examples of this type of solutions are those provided by the main PaaS (Platform as a Service) providers such as Microsoft Azure (Azure, 2022) or Snowflake (SnowFlake, 2022). These centralized platforms simplify in many cases the developments necessary to implement this type of solutions. However, they present a series of drawbacks that prevent or hinder their implementation in real use cases. One of the main disadvantages of these centralized platforms is that they create duplications which makes them difficult to scale. In addition, this type of centralized solutions favours the vendor lock-in, that is, the high dependence on the platform manager. Most companies already have their own platforms and are generally unwilling to implement a new solution due to the cost involved. Therefore, most prefer to reuse the solutions they already have and integrate them with those of third parties using open source and technology-agnostic solutions that do not bind them to any particular provider.

For those reasons, during the last years **decentralized platforms**, also known as **federated**, are imposing themselves as the preferred solution for companies for the development of data sharing platforms (Azure, 2022; SnowFlake, 2022; Constantinides, Henfridsson, & Parker, 2018; De Reuver, Sorensen, & Basole, 2018; Demchenko, De Laat, & Membrey, 2014).

The European Union, promoted by the German and French Governments, has recently defined the technical specification of the GAIA-X architecture. GAIA-X is a federated architecture specifically designed to create a data ecosystem in accordance with European values and standards. The architecture is based on existing standards, and open-source technologies with the aim of creating a federated architecture that competes with the main current cloud platform providers from the US and China and facilitates interoperability and interconnection between its participants allowing the exchange of data and services according to European laws and rights. The GAIA-X architecture is perfectly aligned with the European Data Strategy mentioned above. GAIA-X's intention is to provide businesses with an easy, secure way to exchange high-quality data and services (German Federal Ministry for Economic Affairs and Energy, 2020).

The GAIA-X architecture consists of two clearly differentiated parts: 1) infrastructure ecosystem and 2) data ecosystem:

- On the one hand, the **Infrastructure Ecosystem** focuses on providing or consuming infrastructure services, which in GAIA-X are mainly represented by the asset called Node.
- On the other hand, there is the **Data Ecosystem** where the main asset is the data and services associated with that data. GAIA-X defines a set of federated services that are grouped into four domains:

1. **Identity and trust**: Security is one of the cornerstones of the GAIA-X architecture and therefore a "security by design" approach is used. As part of this approach, the following aspects are defined in detail: Federated Identity Management, Trust Management, Federated Access.
2. **Federated Catalogue (Interoperability)**: GAIA-X provides common protocols and information models based on standards to be able to easily search for the data and services offered by the different actors and to be able to exchange said data and services in a systematic and standardized way.
3. **Sovereign Data Exchange**: The sovereignty of data exchange is guaranteed using corresponding control mechanisms and a concept of global security closely linked to identity and trust.
4. **Compliance**: Establishes organizational solutions to ensure data security and protection. To this end, the following aspects are defined: Relationship between service providers and consumers, Rights and obligations of the participants, Incorporation of new actors and Certification.

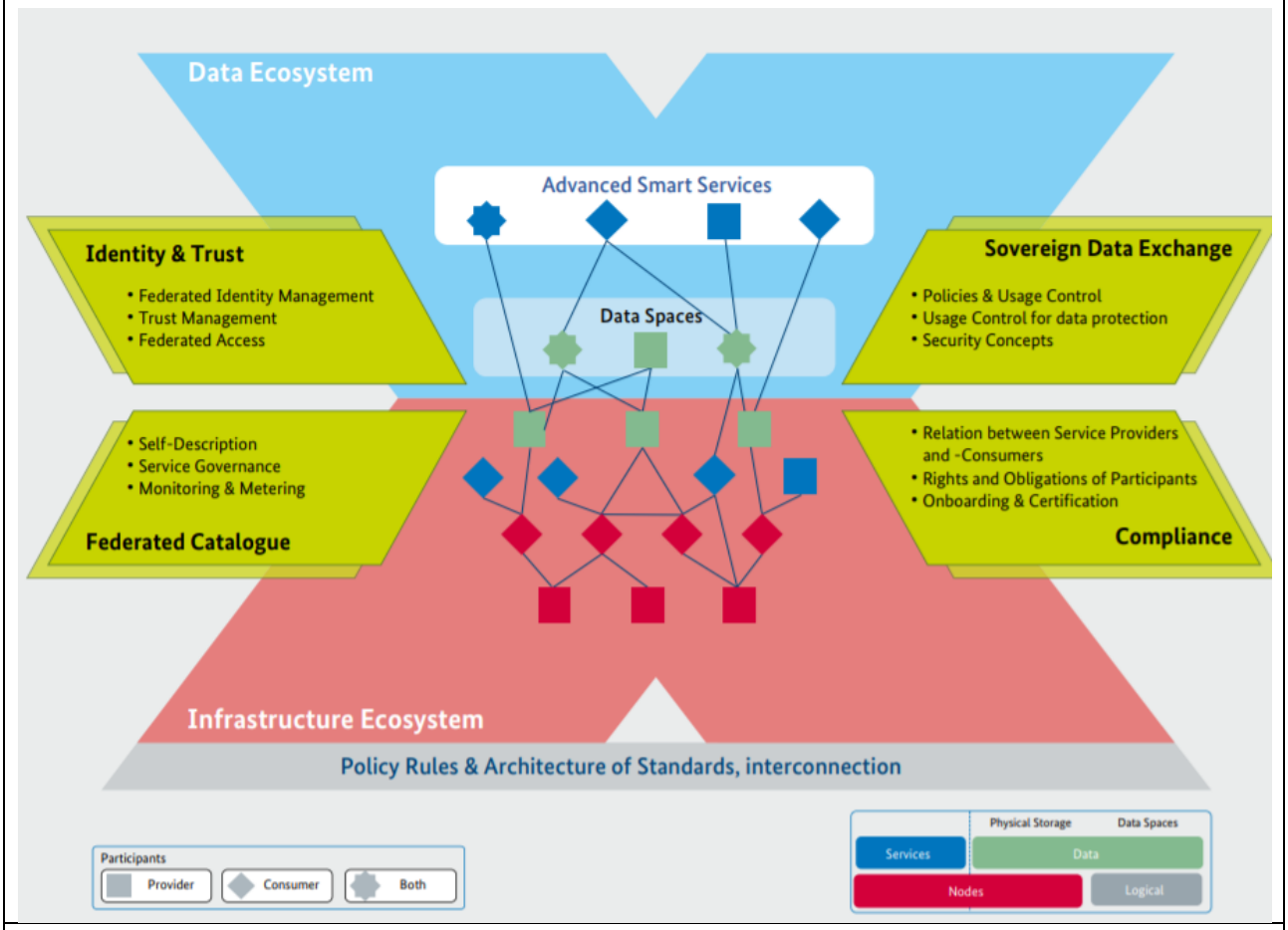


Figure 3.6: High-level architecture Gaia-X

One of the fundamental pillars of Gaia-X's architecture is data governance and sovereignty. That is, the mechanisms that allow to control the use of the data (what data / services I allow to use, who can use them and for what). During the last years, mainly driven by the arrival of data spaces and the "European Data Strategy", multiple studies and possible solutions have been developed to guarantee the governance and sovereignty of the data although none of them has yet managed to have a large-scale adoption (Al-Ruihe, E, & Hameed, 2018; Alhassan, Sammon, & Daly, 2016).

In this field, one of the most established solutions is the solution offered by IDSA (International Data Spaces Association). IDSA is a non-profit organization that brings together more than 120 organizations (mainly companies and technology centres) and whose objective is to develop, certify and maintain a global standard

for international data interfaces and spaces, called IDS (International Data Spaces), as well as to promote technologies and business models related to the data economy. To this end, IDSA has developed a high-level decentralized (federated) reference architecture consisting of five layers (Business, Functional, Process, Information and System) that brings together the concerns and points of view of various stakeholders at different levels of granularity. In addition, this reference architecture is complemented by the specifications of the different components that make up this architecture (Spaces, 2019):

- **Connector**: Provide standardized connectivity in the IDS ecosystem. Responsible for connectivity and control of data use. They also allow the use of secure apps (installable on the connector itself). There are connectors in the data provider and in the data consumer.
- **Identity provider**. Manages the identity of the different stakeholders, their permissions and obligations.
- **Vocabulary Provider**. Manages and offers vocabularies to understand data and facilitate interoperability.
- **Broker**. Allows to search for data and consult the definition, quality, structure and other attributes of the same.
- **App Store**. It allows to search for applications that process raw data (transformation, aggregation, data analysis).
- **Clearing House**. Records data transactions and associated services between data providers and consumers.

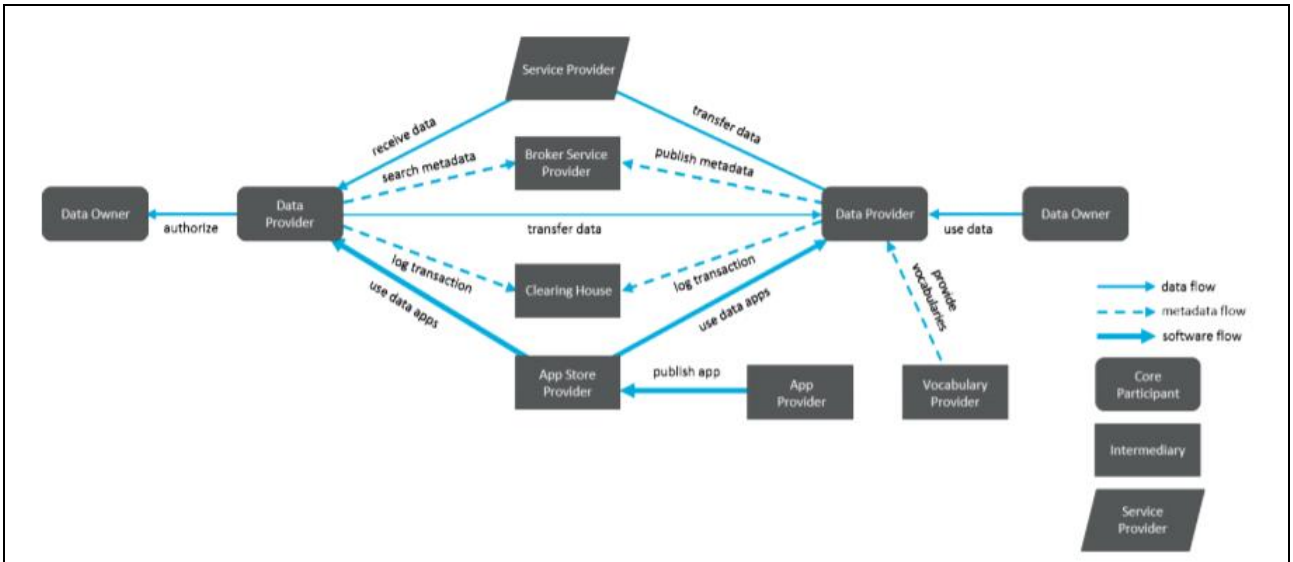


Figure 3.7: IDS Schema GAIA-X

IDS Schema. A fundamental pillar of the GAIA-X architecture.

IDS is a fundamental pillar of the GAIA-X architecture when it comes to data sovereignty. As a leader in this field, IDSA has contributed its knowledge to GAIA-X since its inception and is a founding member of GAIA-X AISBL, the non-profit association that promotes the GAIA-X initiative. The position paper written by Otto et al. analyses in detail the complementarity of both architectures (B, A, A, GAIA-X, & IDS, 2021).

In conclusion, in the literature there are many examples where different solutions for data governance have been investigated. As a summary, the BRIDGE H2020 initiative (EU, 2020), funded by the European Commission and which brings together the H2020 projects of Smart Grids, energy storage, islands, and digitalization, has developed a common framework that contains the main standards related to data governance at different levels:

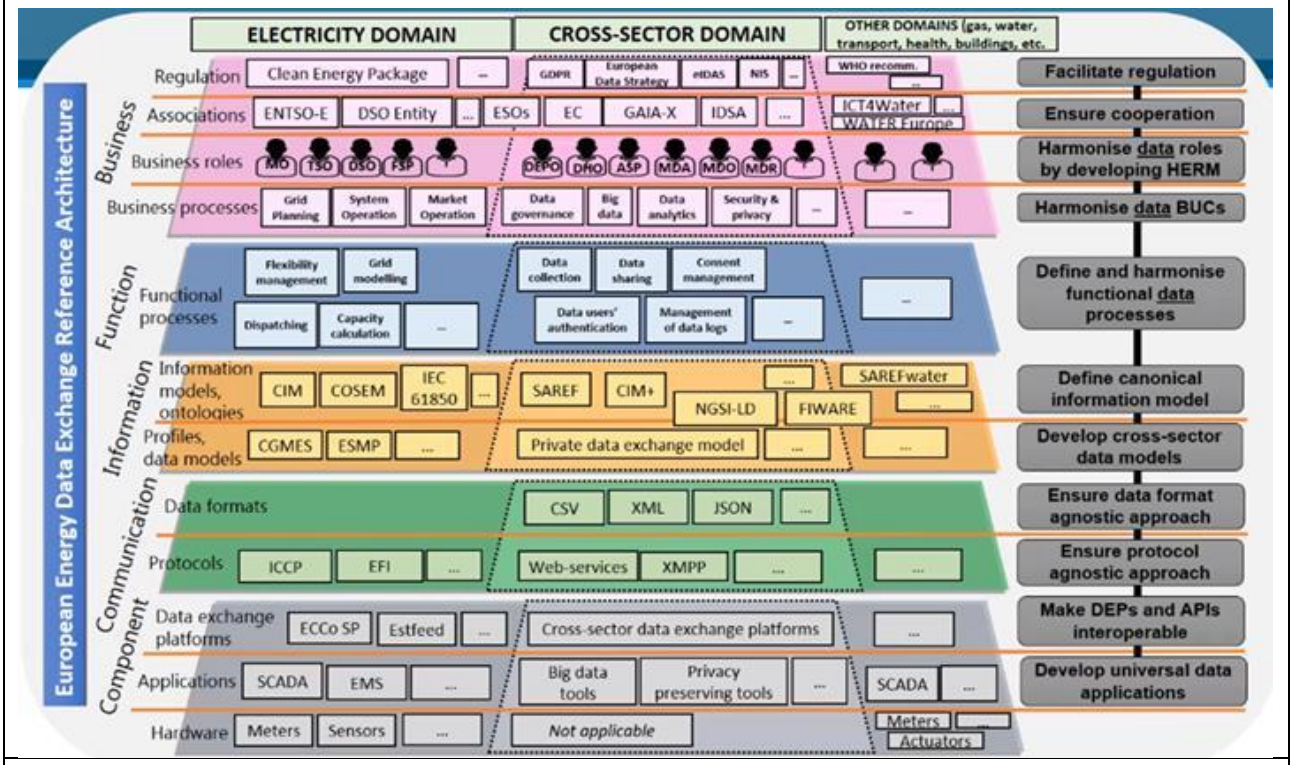


Figure 3.8: Bridge H2020 Framework

Common data governance framework as defined by BRIDGE H2020

3.3 Database

There are 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 so any detailed long-term recommendations could quickly become obsolete, there are several needs that can be defined that must be guaranteed such as response speeds, information consistency, robustness, security, and availability. This section also aims to help those in need of a data storage solution to make the best choice for them and their data. To note; to cover all possible options this section would probably require its own report, therefore, where more information is available at the want of the read, resources are left in the text.

3.3.1 Large datasets

These next sections address the problem of storing large amounts of data.

3.3.1.1 Ideal

The goal of the following sections is to address the problems of storing large amount of veloce data. This is also known as the 3V/5V of the big data (for more information see - https://www.linkedin.com/pulse/3-vs-7-whats-value-big-data-rajiv-maheshwari/:

- Volume: large amount of data
- Velocity: data arrives at a high rate
- Variety (it will be less considered here as we will not focus much on photos and/or text data)

Sometimes, the two other “Vs” are added: one for “Veracity” which is about the quality of the data, and another for “Value” which is about the potential of the data to find insights.

Insights on various data storage options:

1. Data lake – file-based data store where the data structure is optional
    - A very basic example of a data lake is a set of multiple CSV files of measured PV production ingested regularly and stored in e.g., Amazon S3 bucket. The client has to download the whole file first, understand the structure, the meaning of columns and all the metadata. Only then can the data be queried and used in further analysis.
    - As a next step the description of raw files structure can be externalised by technologies like Amazon Athena (a serverless interactive query service that makes it easy to analyse data directly in Amazon S3 using standard SQL). In this step the metadata can also be defined (site location, sensor type, year, month, etc.) to create partitions over the big ‘table’. This meta information can also be queried by SQL-like language together with data records. Prepared Athena queries can then be used as data sources in analytical dashboards.
2. Data warehouse – database-like store where the structure is mandatory
    - Data warehouse integrates metadata (data catalogue) with the data records typically in the relational database (RDBMS).

In case of treatment of high volume of data, it might be worth considering reducing the raw size with lossless and/or lossy techniques.

Timeseries data are usually very easy to compress (typically between 5x et 10x with zip or similar compression algorithms). Most of the stored data are almost the same from one timestamp to another. Times and values usually evolve slowly and therefore between two timestamps most of bytes to store are almost the same. The common part can be reduced by using some encoding/compression algorithm, like the Huffman coding: https://en.wikipedia.org/wiki/Huffman_coding.

Some timeseries database even goes further:

- delta-delta encoding: store difference and not the actual data
- Simple-8b: use less digits/precision, especially that delta can be very small
- XOR-based compression

Much deeper explanation can be found here:

- https://blog.timescale.com/blog/time-series-compression-algorithms-explained/
- https://blog.acolyer.org/2016/05/03/gorilla-a-fast-scalable-in-memory-time-series-database/

Some data historian also uses some "signal processing" compression by not repeating data points that are in the interpolation of the 2 last data points (given some error margin).

Here is a video and an article explaining OSISoft PI compression:

- https://youtu.be/89hg2mme7S0
- https://www.linkedin.com/pulse/data-engineering-osi-pi-enable-advanced-analytics-ratish-sharma

Sometimes the granularity that is needed in the short and long term is not always the same:

- Finer grain for recent data
- Aggregated data for older data

Some timeseries database propose to automatically save space by automatically aggregating data depending on its age:

- https://docs.timescale.com/timescaledb/latest/how-to-guides/data-retention/data-retention-with-continuous-aggregates/#data-retention-with-continuous-aggregates

For example, the last 6 months of data can be stored at the original granularity, from 6 to 12 months the data is aggregated to 30-minute resolution, and anything older is at 24-hour resolution. Working in tandem with this aggregation is the idea of hot and cold storage offered by micro services such as Azure and AWS– hot storage being rapidly accessible, for the more recent data, whereas cold storage being for the older data that theoretically is not needed as often. The benefit being that the cold storage is significantly cheaper, allowing the storage of significant amounts of rarely used historical data at a lower cost.

3.3.1.2 Read/Write patterns

Before designing/choosing a high-volume database, the read/write patterns need to be understood.

Examples:

- Plotting data: low latency expected, low volume
- Fleet analysis: medium latency with high read performance
- Country/Continent scale metrics storage: high write performance

The choice of the database and their configuration will depend on the expected read/write patterns. We might also need different databases for different patterns; therefore, we might need to duplicate data.

In case you don't have time to build your own benchmark yourself, you can have a look at the Time Series Benchmark Suite (TSBS): https://github.com/timescale/tsbs. Data and queries already exist. It's not always easy to find the results of the benchmark. You should also note that the results depend on when it was run, as all the software have many releases a year, results keep changing.

3.3.1.3 Distributed/replicated database

If the volume of data cannot be handled on one machine and/or there is a need of 24/7 services with 99.9+% of availability, a distributed/replicated database is needed. It's important to understand the CAP theorem when building such architecture:

https://aws.amazon.com/blogs/startups/distributed-data-stores-for-mere-mortals/

Almost all cloud providers have at least SQL and/or NoSQL databases that can scale for billions of records. Some of them also have timeseries oriented SaaS solutions:

- Azure Times-Series Insights
  https://azure.microsoft.com/en-gb/services/time-series-insights/#product-overview
- Amazon Timestream
  https://aws.amazon.com/timestream
- Google Cloud Platform - only recommendation, not a real time-series database
  https://cloud.google.com/bigtable/docs/schema-design-time-series

There are also some OpenSource alternatives like InfluxDB, TimescaleDB, QuestDB, OpenTSDB, Prometheus, Warp10/SenX. These alternatives usually offer a Cloud deployment allowing the use of their services without the need for complicated configuration. This list is long and evolving. Timeseries database are still actively developed and have a strong user community. This “software field” is very active since 2010, they offer evolves fast and new products appear and disappear.

3.3.1.4 Data historians

Time-series databases are popular nowadays but there has been other dedicated software long before, called Data Historians:

- OSISoft PI: https://www.osisoft.com/pi-system

- GE Proficy: https://www.ge.com/digital/applications/proficy-historian
- ABB Historian (also called RTDB or 800xA History...)
- AspenTech's InfoPlus 21
- HoneyWell
- Siemens

The difference with the Time-series database is mainly that are more oriented to industrial needs and support natively some of their protocols (ie: OPC...). More information can be found here: https://www.controleng.com/articles/the-data-historians-history-told/.

3.3.1.5 Minimal setup for medium to large size dataset

For a "not too large" dataset (up to 1B points), a standard SQL database is the best option. Recommendations for how to put:

https://towardsdatascience.com/what-if-i-tell-you-rdbms-can-handle-time-series-data-77a5bb43da06

https://blog.timescale.com/blog/tip-tuesday-february-2020/

With any other system than a database, such as a basic csv file, would result in the loss of an abstraction layer that takes care of data format and the ability to easily request and access data (SQL layer that can be used with ODBC, JDBC or a native interface).

Data can be in long or wide format. Which is the best can depend on several factor. If there are a range of different measurements taken at aligned times (I.e., a measurement of energy production, temperature and current every 5 minutes), a wide approach is perhaps better. This has the benefit of being both easier to read, and more economical in data size when storing the data (will not have to repeat the date three times if had it in the long format). If, however, the timestamps are not aligned, then a long and narrow table is perhaps better (as shown below).

Wide table

Timestamp	**GHI**	**CloudCover**	**Temperature**	**...**
2020-03-20T01:30:00+00:00	523	0.74	12.5	...

Narrow and Long Table

Timestamp	**Metric**	**Value**
2020-03-20T01:34:00+00:00	GHI	523
2020-03-20T01:45:00+00:00	CloudCover	0.74
2020-03-20T01:52:00+00:00	Temperature	12.5
2020-03-20T02:08:00+00:00	GHI	656
2020-03-20T02:12:00+00:00	CloudCover	0.78
...	...	...

3.3.2 Small datasets

3.3.2.1 CSV

For smaller datasets, up to a few millions of records, it might be easier to use files even if it is not the recommended solution.

The main issue with text file formats, such as CSV, is that some datatypes can be badly interpreted:

- Dates: is it local or not? Does this field represent a day or a month…?
- Strings: it gets complicated when you have a separator in the string field, and many other cases…
- Numbers: what is the decimal separator? Scientific notation might pop up too…
- Complex types such as arrays

Despite this, for many small dataset purposes, a simple CSV file is the best solution for storing data and sharing data.

The practical and classic solution is to keep together column values 'indexed' by the same timestamp in a table-like structure, e.g., CSV file. In this case metadata must be stored separately, e.g., in the file header or in separate database tables. However, the different data structures can be required for certain data storages (e.g., cloud-based data lake storage). In these storages the metadata (sensor name, site location, etc.) are an integral part of every record and represent dimensions by which measured values can be queried.

The usual solution for the monitoring system is to generate daily files of data (starting from midnight 00:00:00 to 23:59:59), but other configurations are possible, such as monthly files if the recording interval is not low so as not to make too big files.

The monitoring system should generate complete files with all data available, but then extraction can be done if limited files are required for a given purpose. Or, on the contrary, new columns can be added to the files with calculated data (ex: yields) or flags.

It is also possible to make "minimum" files, with datetime, energy and irradiation for instance, at a few-minute time step to get a near real-time view on the plant operation. Another way to make short files, in order to save space on hard disks, is to limit the recording time to daylight time or even when irradiance is higher than 20 W/m² (IEC 61724-1).

Another type of data is often well appreciated in the data analysis is the "log" file (standing for "logistic"), which lists all the events that have occurred in the plant and the maintenance actions. Each event or action should be time-stamped as well.

3.3.2.2 Better file formats

There are 2 categories of file format: human and not human readable. Performance wise it is usually better to use not human readable format as there are more compact and are sometimes very well optimized for some use cases (columnar storage, hierarchical data, sparse matrix…).

Not all the data formats are covered, for more information see:
https://en.wikipedia.org/wiki/Comparison_of_data-serialization_formats

Human readable

The two most common human readable file format are JSON and XML. From 2010, most people moved from XML to JSON as it is easier to use and have less constraints. For strong data validation, XML is a better choice as it has an official schema description and there are lots of tools available used to validate an XML file.

Schema descriptions can be found here:

- JSON (nothing official for now): http://json-schema.org/
- XML Schema (official): https://en.wikipedia.org/wiki/XML_Schema_(W3C)

Not human readable

The main advantage of non-human readable format is that the data types are handled properly if the code also uses the proper types. Many mistakes are then avoided regarding type conversion.

There are many not human readable files format targeted to different usages. The two most useful given our context:

- Apache PARQUET: recently the PARQUET columnar storage has had a lot of tractions. It is used in the Big Data eco system widely. It is fast to read data and allow parallelism on the data too. More details here: https://en.wikipedia.org/wiki/Apache_Parquet. There is an effort to move forward with the Apache Arrow project: https://en.wikipedia.org/wiki/Apache_Arrow
- HDF / NetCDF: there are older file format, initially designed for super computing and complex type as arrays. They have very good performance.
  https://en.wikipedia.org/wiki/Hierarchical_Data_Format

3.4 Fog/edge computing (edge + cloud)

Though somewhat outside the remit of this report, it is still worth highlighting the use of cloud computing for handling data monitoring and storage. Fog/edge computing relate to an architecture where the Cloud and the IoT devices are not the only ones in charge, but an edge device is introduced between the two (some IoT devices with enough computing/storage capacities can be considered as an edge device too). The advantages having an edge are mainly:

- The edge is usually physically relatively closed to the IoT devices and have a good connectivity to the IoT devices: the connection is supposed to be more qualitative in term of availability, latency, and bandwidth than the Cloud. Therefore, the quality of service should be better on the IoT side in case of Cloud connectivity degradation/disconnection
- In the case the data needed to “operate” the IoT device are available locally in the edge, and the data in the Cloud are not needed. Having the intelligence or (most of the intelligence) in the edge instead of in the Cloud limits the amount of data needed to be transferred to the Cloud. The edge can aggregate and/or build statistics on the data and send them to the Cloud. In that way we can use the Cloud to build reports to have a fleet vision, make some adjustments to the IoT devices parameters / behaviours using the statistics sent…

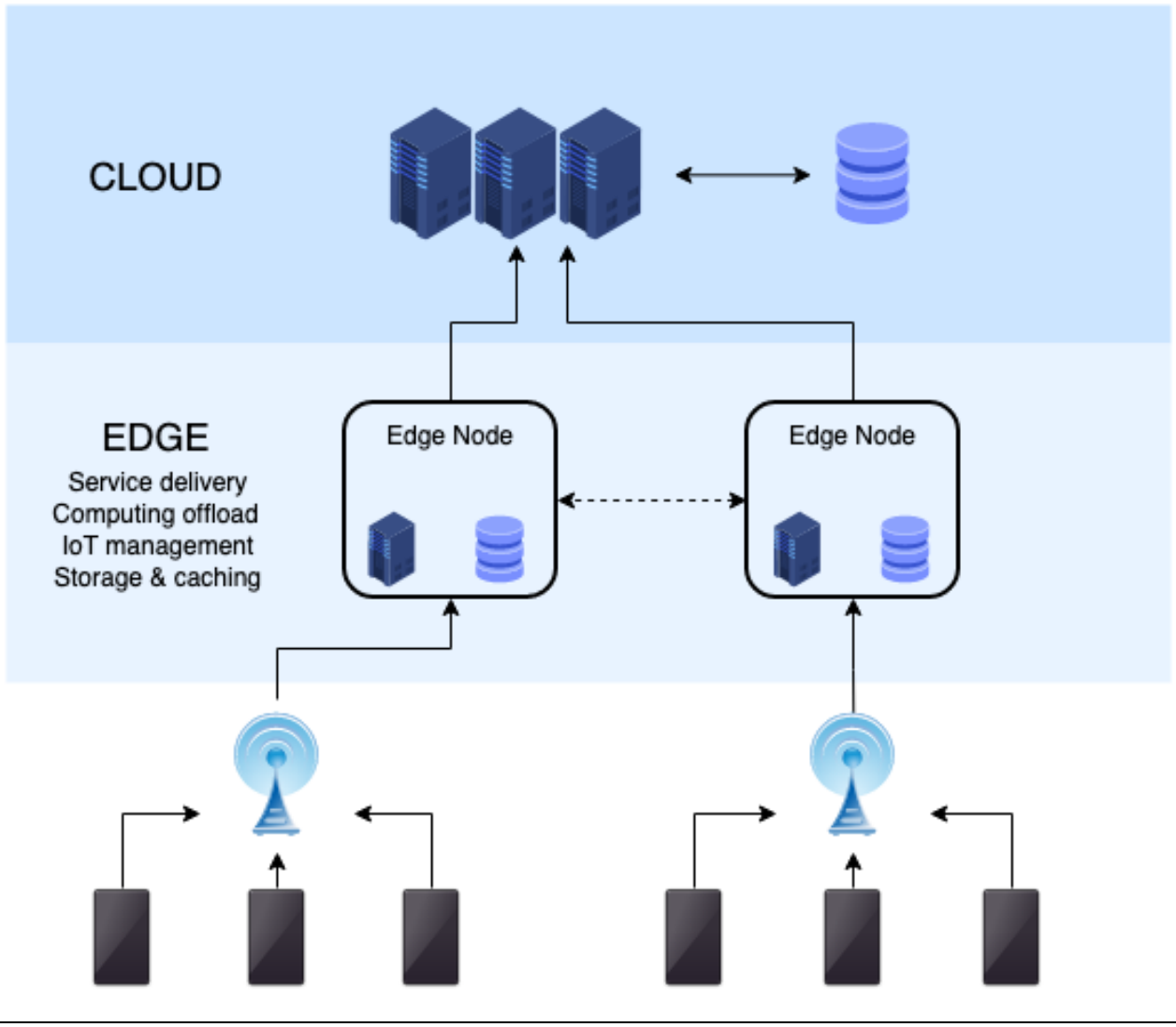


Figure 3.9: Edge Computing Schema

Schema taken from https://en.wikipedia.org/wiki/Edge_computing showing the architecture of cloud storage.

There are many commercial offers for edge/fog computing:

- Azure IoT Edge: https://azure.microsoft.com/en-us/services/iot-edge/#iotedge-overview
- Amazon Greengrass: https://aws.amazon.com/greengrass/
- Google Edge TPU: https://cloud.google.com/edge-tpu
- HPE IoT Edge: https://www.hpe.com/us/en/solutions/edge.html

3.5 Transfer Protocols

Data needs to be shared, be it 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. In order to be able to do this, data transfer solutions and protocols need to exist to enable the simple, as quick as possible, 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.

3.5.1 JDBC/ODBC

For databases (and timeseries databases): a SQL layer is very desirable as many tools can be used to access and visualize the data.

There are two very common protocols that are supported by almost all databases:

- ODBC: usually a bit more oriented towards Microsoft systems, but well ported on other platforms recently
- JDBC: more oriented to Java application

There are JDBC->ODBC and ODBC->JDBC bridges in case the chosen database or coding language does not support both.

Moreover, the advantage of a database is that different rights can be granted to different people (read/write/update…). It is also possible to limit what users can see, or only show them aggregates of one or multiple tables, using views. More details on views here: https://en.wikipedia.org/wiki/View_(SQL).

The protocol will not bound the amount of data to be transferred and queries are easily to parallelize. The bottleneck is usually the database itself.

3.5.2 Messaging/IoT systems

Another, more modern, way to exchange data is to use a message queuing system so that is possible to send and receive data in a more streaming oriented manner: https://en.wikipedia.org/wiki/Message_queue. We will specifically present the publisher/subscriber (also called pub/sub). The advantage of such system is that is allowing better scalability and looser coupling between application producing and consuming the data. Transferring terabytes of data daily using this system is a very common use case.

The most known cloud providers has such services:

- Azure IoT Hub: https://azure.microsoft.com/en-gb/overview/iot/
- Amazon AWS IoT: https://aws.amazon.com/iot/
- Google IoT: https://cloud.google.com/solutions/iot

It is also very common in the industrial IoT world (we cannot list them all as there are 10s of them):

- ABB Ability
- GE Predix
- C3 IoT
- IBM Watson
- Hitachi's Lumad
- Siemens Mindsphere

All those platforms are targeted towards what is called the “Industry 4.0”. More details here: https://en.wikipedia.org/wiki/Fourth_Industrial_Revolution

In essence, the Fourth Industrial Revolution is the trend towards automation and data exchange in manufacturing technologies and processes which include cyber-physical systems (CPS), IoT, industrial internet of things, cloud computing cognitive computing and artificial intelligence.

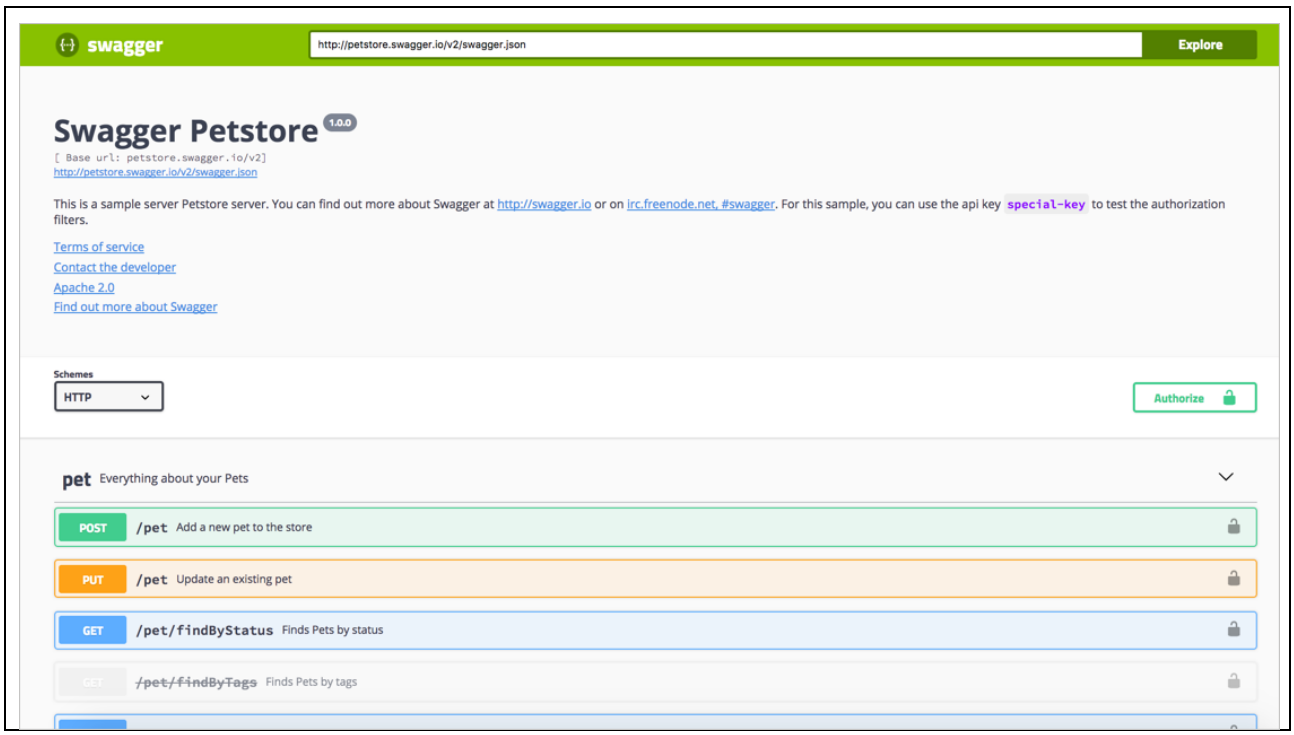
3.5.3 SFTP

It is not recommended to use FTP/SFTP servers as using a messaging system should be preferred to transfer data. If it is unavoidable (it makes sense for a one-time (potentially massive) data transfer), the best option is to use a SFTP server (FTP) which is properly secured contrary to a standard FTP. The best is to use public and private SSH keys instead of passwords to secure the connection, and to generate a key for each user/application. More details on how to do that here: https://www.ssh.com/academy/ssh/keygen.

3.5.4 API (HTTPS REST)

Nowadays most of the applications exchange data using REST APIs. It is based on the HTTP protocol: https://en.wikipedia.org/wiki/Representational_state_transfer. It is fairly easy to use and to implement client and server side for an application, as many tools and programming languages have ready to use implementation / libraries. Many tutorials covering the different programming languages are available.

There are also many simple ways to share at the same time both the data specification and documentation using tools like SWAGGER (https://swagger.io/):



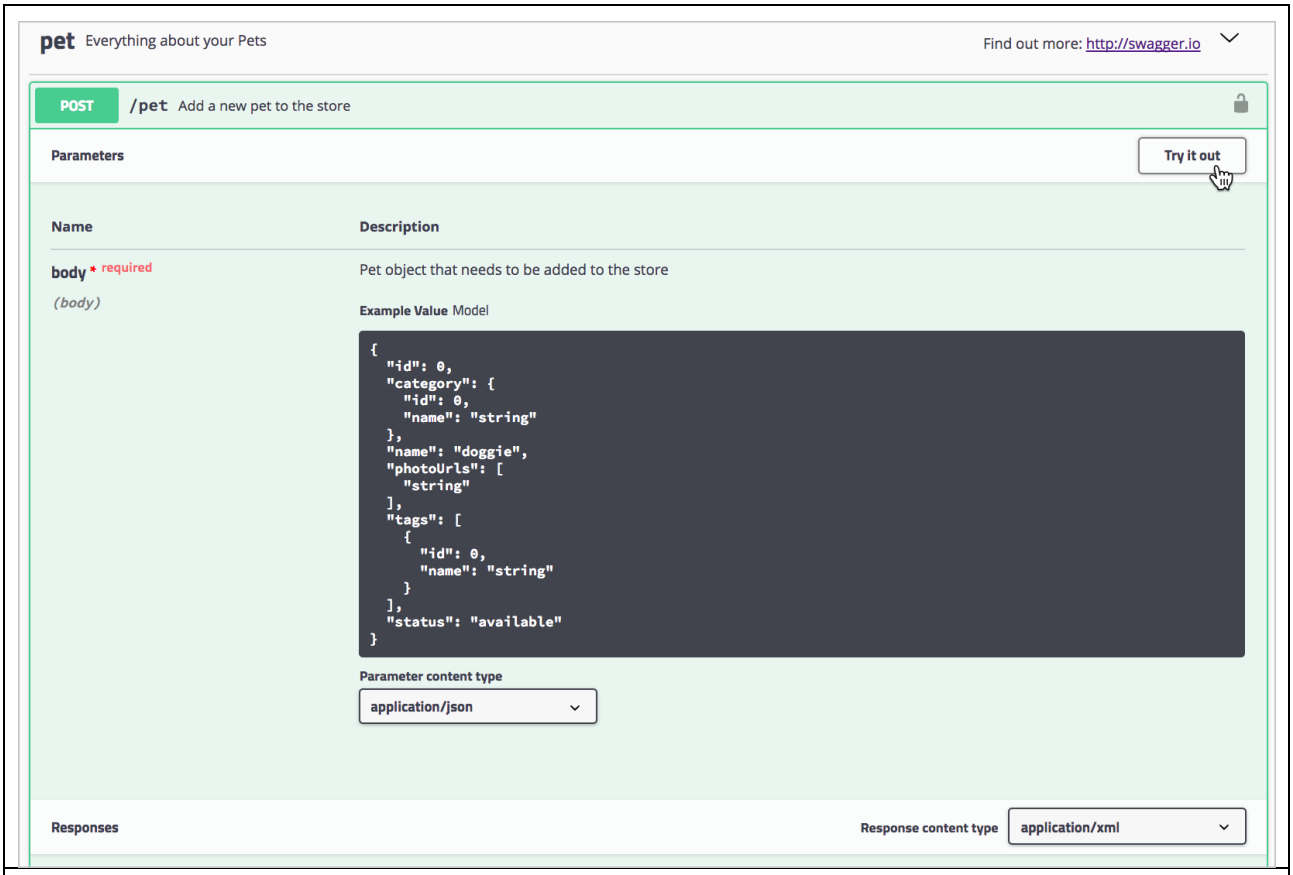


Figure 3.10: Swagger API method documentation

It can also automatically create the code to be used on the client side:

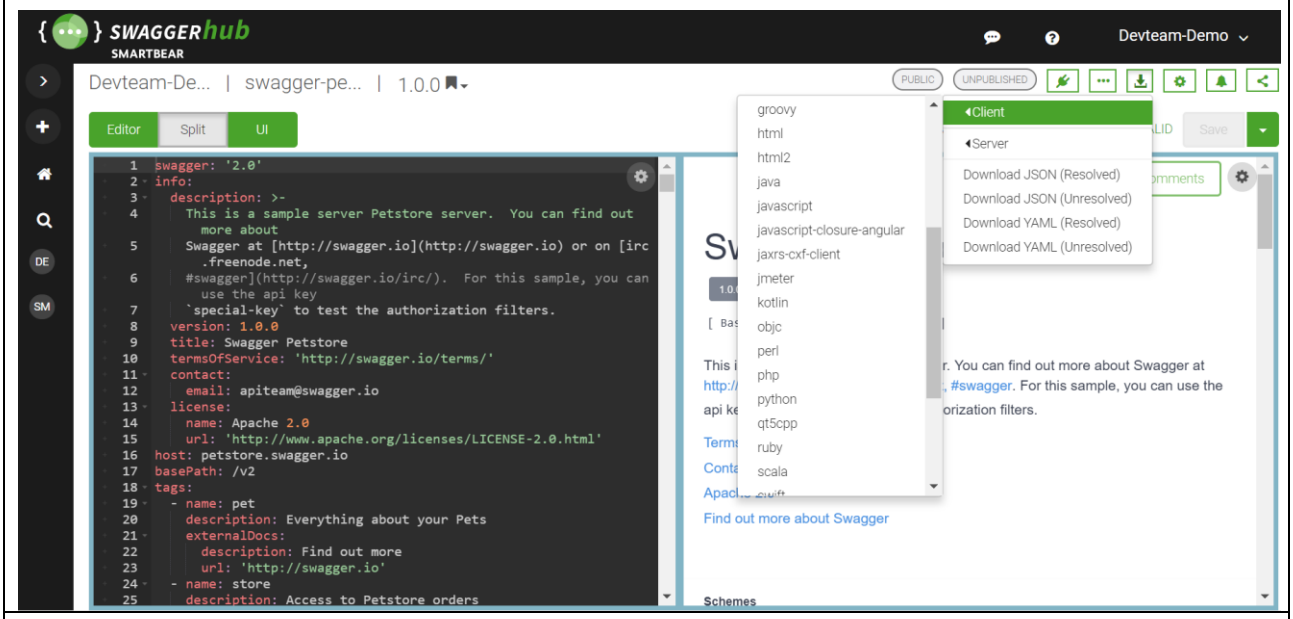


Figure 3.11: Swagger automatic code generation example

3.6 Data privacy, sovereignty, security 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 GPRD), 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.

3.6.1 Roles

There are a variety of different roles and actors within the photovoltaic community. In function of the role, function or legal status of a data collect or transfer/sharing or transformation, the rules and what to consider may differ.

Examples of the stakeholder categories identified as:

- **commercial partners or service suppliers** are companies providing (1) Operations and Maintenance (O&M), (2) Asset managers, (3) Forecasting data providers, (4) Monitoring data providers, (5) Consultancies.
- **regulators or governmental authorities** there are (6) regulatory bodies, (7) Distribution System Operators (DSO), (8) Transmission System Operators.
- **financial institutions** we have considered (9) Investors and funds, (10) Banks and financial institutions providing loans for PV projects, (11) Insurance companies.
- finally, in the **PV Energy management and trading** there are (12) Flexibility providers and (13) Energy traders.

3.6.2 Definitions

For use as a reference, the following are a list of definitions of terms used when dealing with data ownership, privacy and sovereignty:

- '**personal data**' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
- '**confidential data'** means any data that is supposed to remain secret i.e. not known publicly. Personal data can be confidential; however, it is important to make the difference. Examples of confidential data are financial records or intellectual property contracts.
- '**processing**' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;
- '**restriction of processing**' means the marking of stored personal data with the aim of limiting their processing in the future;

- '**profiling**' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;
- '**pseudonymisation**' means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person;
- '**filing system**' means any structured set of personal data which are accessible according to specific criteria, whether centralised, decentralised or dispersed on a functional or geographical basis;
- '**controller**' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;
- '**processor**' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller;
- '**recipient**' means a natural or legal person, public authority, agency or another body, to which the personal data are disclosed, whether a third party or not. However, public authorities which may receive personal data in the framework of a particular inquiry in accordance with Union or Member State law shall not be regarded as recipients; the processing of those data by those public authorities shall be in compliance with the applicable data protection rules according to the purposes of the processing;
- '**third party**' means a natural or legal person, public authority, agency or body other than the data subject, controller, processor and persons who, under the direct authority of the controller or processor, are authorised to process personal data;
- '**consent**' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;
- '**personal data breach**' means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed.

3.6.3 Data Types

The following table is taken from the *D7.3: Assessment and recommendations on legal aspects, policy, confidentiality, data privacy* and lists various options for data classification relevant in the PV sector depending on the structure, confidentiality, time frame, and nature of the data

Table 3.2: Data Classification

Classified by		**Definition**
Structure	**Data vs Database**	**Data** is the primary data, not treated or manipulated in any way. Sometimes mentioned as a digital object. **A database** can be considered as structured data. It is the system where the information is collected.

	Structured vs Unstructured data	**Structured data** is formatted and organized in a pre-defined way so that processing and analysis can be applied. **Unstructured data** is not organised or defined before being sent.
Confidentiality	**Confidential vs Public data**	**Confidential data** refers to personal information which is shared in confidence with another party. Confidentiality agreements are often implicit. **Private** data may be read by the users with access to that data library, while **public data** is accessible by all users.
	Pseudonymized vs Anonymized data	**Anonymization**: when it is not possible to restore the original information. **Pseudonymization** replaces sensitive data, subject can still be identified through indirect or additional information
Time frame	**Real-time, Delayed and Reference data**	**Real-time** (chargeable) data is delivered virtually instantly with its creation. **Delayed data** (non-chargeable) varies between data providers. **Reference data** includes historical or non-real time information.
Nature of the data	**Personal** vs **non personal** **Financial** – e.g. CAPEX, OPEX, loans… **Technical** data – e.g. PV production, weather forecast, data availability **Time series** data- e.g; spot price, generation or outage time series **Geospacial** data **Environmental** e.g. impact on biodiversity **Legal** – e.g. data required for permitting	

3.6.4 Data Privacy and Data Sovereignty

Data privacy concerns the proper handling of sensitive data including, notably, personal data but also confidential data, such as certain financial data and intellectual property data, to meet regulatory requirements as well as protecting the confidentiality and immutability of the data. Personal data means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

When considering data privacy there are several things to consider:

- What data is needed vs What can be shared?
    - How accurate do coordinates need to be?
    - What granularity of the data is required?
    - Can anything be inferred about the residents' lives if residential installation?
    - Is plant financial information inferable or available?
- GDPR (and future British equivalent)
    - Defines rules applicable for personal data for EU and EU residents
    - Aims to protect fundamental rights and freedom of natural persons and their right to protection of personal data
- Personal vs Confidential Data

3.6.4.1 Data Anonymisation

One of the ways to solve data privacy issues is through the use of Data anonymisation. There are many techniques to anonymize data in order to be able to share them without disclosing personal information. GDPR requires that personal data need to be stored after being anonymized or pseudonymized. You can find some overall information in Wikipedia: https://en.wikipedia.org/wiki/Data_anonymization and a survey [ref] those techniques.

We will focus on 3 techniques related to the SERENDI-PV project:

- Pseudonymization: to share data within participants in the project
- K-anonymity / differential privacy: if we need to share data outside the project
- Time-series anonymization: to avoid leaking too much on people usages or even identify someone

Pseudonymization

Pseudonymization is the simplest way to anonymize data as names, cities, birthdate, postal codes etc. It means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person. There are different techniques to pseudonymize data, like directory replacement, scrambling, or masking. In general, pseudonymization it is done by replacing the values of the field with another value using a hashing algorithm. Dictionary attacks can be avoided by the addition of some "constant" to the value before hashing (called salt) – this salt should not be shared.

The Dataiku documentation give some more details about all this:

https://doc.dataiku.com/dss/latest/preparation/processors/column-pseudonymization.html

As does Wikipedia:

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

K-anonymity

K-anonymity (Samarati & Sweeney, 1998) is not a process by itself, but a property of a dataset. The dataset is considered to be at least k-anonymize if it is not possible to distinguish a person from k-1 other people.

There are two common methods for achieving k-anonymity:

- **Suppression**: certain values of the attributes in a column which have too low cardinality are replaced by a new value or the column can be even removed.
- **Generalization**: individual values of attributes are replaced with a broader category or a range.

https://en.wikipedia.org/wiki/K-anonymity

Differential privacy

Differential privacy is used to publicly shared statistics on a dataset without sharing detailed information. It is usually achieved by adding noise to the data so that we cannot distinguish individuals within the dataset. It was developed because it was proven that using "reconstruction attacks", a larger part of the original private data could be recovered (see also: https://en.wikipedia.org/wiki/Reconstruction_attack).

More details can be found in this Wikipedia article: https://en.wikipedia.org/wiki/Differential_privacy and in (Dwork & McSherry, 2017)

Time-series anonymization

Time-series can also contain private data. For example, your energy consumption could show your habits/household characteristic: such as waking up and bedtime, if you were home at a given day/week/weekend, the kind of home appliances you have… A long enough time-series can easily be a unique identifier of a household. Another example would be with solar production data, you could infer the longitude and latitude of a building.

There is also an area of research to find appliances from a load curve that could reveal some element of privacy of the household. Such process analyses changes in the voltage and current going into a house and deduce what appliances are used. This is usually called Non-Intrusive Load Monitoring (NILM) – there are workshops (http://nilmworkshop.org/) and open-source libraries (https://github.com/nilmtk/nilmtk) related to this area.

There are many ways to improve time-series privacy:

- direct perturbation:
  - time: aggregating data to a lower resolution. For example, moving from 5-minutes data to hourly or daily data
  - precision: limit the resolution of the measures (for example: from Wh to kWh), or replace them with a categorical/range value (or as a letter in the SAX algorithm: Symbolic Aggregate approximation)
  - transformation: using Fourier transformation or wavelet to keep important parts of the information and add some noise during reconstruction
  - …
- swapping or concealing: use part of the data from other individuals that are closed to each other to create a new time-series that would be probable but would not make the individual identifiable

3.6.5 Distribution and Intellectual Property

With the transfer, exchange, buying and selling of data, comes the associated rights. It is important that in any data transaction, both parties are clear as to what data is being received and what can and, importantly, cannot be done with that data.

Distribution rights and intellectual property rights are very similar in that they concern the ownership of data or more technically correct a dataset from two different points of view: that of the owner - intellectual property, and that of a secondary party - distribution rights. The distribution rights represent a legal agreement dictating what that secondary party can do with, i.e., sell or use in a particular way and where. Intellectual property data laws however concern who actually owns the data/dataset/database.

The following table taken from the *D7.3: Assessment and recommendations on legal aspects, policy, confidentiality, data privacy* summarises the different types of agreements within a data sharing context.

Table 3.3: Data Sharing Agreements

Type of agreement	Description	Provisions
License agreements. Such as End User Licence agreements (EULA)	Legal contracts between two parties: an owner (licensor) and a second party (licensee). The owner gives official permission to the licensee to do, use or own something (a software, a brand, a patented technology, or the ability to produce and sell goods) by the licensor. So,	• Nature of the agreement, • a Copyright or IP right: clarifying ownership • Limitation of liability clauses • Disclaimers • Governing Law

	a licence agreement grants the licensee the ability to use the IP owned by the licensor. They are commonly used to commercialize IP.	• Right to terminate the agreement • Authorized use • Unauthorized use Other aspects to consider are: third parties, right to modifications, right to sell,
Terms of Use / ToU, Terms and Conditions T&C, Terms of Service ToS, User Agreements, Terms of User Agreements, acceptable Use Policy	These are legally binding agreements between a **service provider** and a person who wants to use that service. The person must agree to abide by the terms of service in order to use the offered service. Terms of service can also be merely a disclaimer, especially regarding the use of websites. Vague language and lengthy sentences used in the terms of use have brought concerns on customer privacy and raised public awareness in many ways.	• Governing law: jurisdiction • Disclaimers • Liability limitation • Rules of Account Termination • Permitted and Restricted Use (including user behaviour/guidelines) • How to register for an Account • Need to specify a right to terminate the owner's services to a specify user and not just the license of the software.
Data sharing agreement	Formal contract clearly documenting data being shared and its uses. It protects the agency providing the data, ensuring that the data will not be misused and it prevents miscommunication on the part of the provider of the data and the agency receiving the data. Before any data is shared, both: provider and receiver discuss data-sharing and data-use issues and come to a collaborative understanding, documented in a data-sharing agreement.	• Period of agreement • Intended use of the data • Constraints on use of the data • Confidentiality • Security • Methods for data sharing • Financial costs for data sharing
Non-Disclosure Agreements (NDA)	Legally binding contract establishing a confidential relationship between the parties involved, to protect information required to do business.	• Parties involved • Definition of confidential information • Disclosure period • Authorized use, • Terms for disclosure • Exclusions • Legal provisions • ToU, or T&C • Law governing the parties

Other types of data sharing agreements include the Standard Software License sharing agreements, Service Level agreements (SLA)

Aspects to consider when defining a **Data Sharing** agreement:

1. Period of agreement
2. Intended use of the data
3. Constraints on use of the data
4. Data confidentiality
5. Data security
6. Methods of data-sharing
7. Financial costs of data-sharing

More information on distribution, IP and contracts between parties can be found in the D 7.3.

3.6.5.1 SERENDI-PV Collaborative Platform

The creation of SERENDI-PV collaborative platform and its operation will pertain more areas and regimes of Intellectual Property and it will be necessary to properly cover the respective regimes, whether on internal basis among the Consortium members and its project partners or externally with respect to the SERENDI-PV users:

1. The first principal area is the one of the creators of the tools, modelling systems and/or databases to be provided on SERENDI-PV which is protected by IP rights and some particularities might need to be discussed and outlined according to the type of creation, origin, and form of collaboration with the concrete authors, co-authorships, etc. SERENDI-PV Consortium should be able to hold/manage all necessary exploitation rights for running the SERENDI-PV collaborative platform and for distributing the respective usage rights or licenses, as the case may be. Agreement/s governing the rules of these creations among the Consortium members should be put in place. Understanding on data terms should also be a part of it.
2. The second principal area is the one concerning the SERENDI-PV collaborative platform´s users and it will be necessary to establish the mode (modes) under which the users will be able to use the main tools (databases, software) and data offered within these tools. Several types of end user licenses and terms of use are available. The final choice should be the results of Consortium´s agreement on the scope of rights, freedoms and limitation of the SERENDI-PV´s users.
3. Other areas to discuss and/or regulate: users´ content (users' data…), access to third parties´ scientific articles or services (if provided), open access policy to other resources (extent, areas to include), disclaimers of SERENDI-PV (in general, on data, on different tools, as the case may be).

4 CONCLUSIONS

This report has discussed and suggested a variety of recommendations for use within the photovoltaic domain. The implementation of these ideas during the SERENDI-PV project, feedback and experience will then be reported in the follow on to this task, task 7.2 in work package 7.

This report has covered several “data” themes within the photovoltaic domain:

- Data Collection
- Data Format
- Database
- Transfer protocols
- Data Privacy and Anonymisation
- Distribution and IP rules

It is difficult to define a set of protocols and one-size-fits-all data rules within the PV as (as shown by figure 2.1) there are a huge number of different actors, different data use cases, data requirements and this will only continue in the future with the advancement of technology, ideas and the further integration of PV into electricity grids across the globe. However, this document tries to act as a “go-to” guide to try and steer the community towards a general standardisation, proposing what current experts in the field consider the “best” option for, for example, the storage of large datasets, the anonymisation of data or the nomenclature of data terms to be used in datasets.

These ideas will next be implemented, given feedback to, and improved upon in the natural follow on of task 1.4, task 7.2.

5 BIBLIOGRAPHY

- Dwork, C., & McSherry, F. (2017). Calibrating Noise to Sensitivity in Private Data Analysis. *Journal of Privacy and Confidentiality, 7*(3), 17-51.
- Samarati, P., & Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Oakland: S&P.
- Accenture. (2018). *Value of Data. The Dawns of the Data Marketplace.* Retrieved 2022, from http://www.accenture.com/t20180904T113809Z_w_/us-en/_acnmedia/PDF-85/Accenture-Western-Digital-Value-of-Data-of-the-Data-Marketplace.pdf
- Otto, B., Lis, D., & Cirullies, J. (2019). *Data Ecosystems: Conceptual Foundations Constituents and Recommendations for Action.* Fraunhofer: Fraunhofer ISST.
- German Federal Ministry for Economic Affairs and Energy. (2020, June). *GAIA-X:Technical Architecture, Release.* Retrieved 2022, from https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/gaia-x-technical-architecture.pdf?__blob=publicationFile&v=5
- Azure. (2022). *Azure* . Retrieved from https://azure.microsoft.com/es-es/
- SnowFlake. (2022). *SnowFlake*. Retrieved from https://www.snowflake.com/?lang=es
- Constantinides, P., Henfridsson, O., & Parker, G. G. (2018). Introduction – Platforms and infrastructures in the digital age. *Information Systems Research, 29*(2).
- De Reuver, M., Sorensen, C., & Basole, R. (2018). The digital platform: a research agenda. *Journal of Information Technology, 33*(2).
- Demchenko, Y., De Laat, C., & Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. Minneapolis: International Conference on Collaboration Technologies and Systems .
- Al-Ruihe, M., E, B., & Hameed, K. (2018). A systematic literature review of data governance and cloud data governance. *Personal and Ubiquitous Computing*, 1-21.
- Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: an analysis of the literature. *Journal of Decision Systems, 25*(1), 64-75.
- Spaces, I. D. (2019). *IDS Reference Architecture Model v3.0*. Retrieved 2022, from https://internationaldataspaces.org/publications/ids-ram/
- B, O., A, R., A, E., GAIA-X, & IDS. (2021). Position Paper v1.0. *International Data Spaces Association*.
- EU. (2020). *Bridge H2020*. Retrieved 2022, from https://www.h2020-bridge.eu/
- Thema Consulting Group. (2017). *Data exchange in electric power systems: European State of Play and Perspectives.* ENTSO-E.
- Gueymard, C. A. (2004). The sun's total and spectral irradiance for solar energy applications and solar radiation models. *Solar Energy*, 423-453.