Development of a big data bank for PV monitoring data, analysis and simulation in COST Action 'PEARL PV'

Angele Reinders[a], Fjodor van Slooten[a], David Moser[b], Wilfried Van Sark[c], Gernot Oreski[d], Bettina Ottersboeck[d], Nicola Pearsall[e], Mirjana Devetaković[f], Jonathan Leloux[g], Dijana Capeska Bogatinoska[h], Christian Braun[i], Anne Gerd Imenes[j], Anton Driesse[k]

a) ARISE, University of Twente, Enschede, 7500AE, The Netherlands, b) Institute for Renewable Energy, EURAC, Bolzano, 39100, Italy, c) Copernicus Institute of Sustainable Development, Utrecht University, Utrecht, 3584 CB, The Netherlands, d) Polymer Competence Center Leoben GmbH, Leoben, A-8700, Austria, e) NPAG, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK, f) University of Belgrade, Faculty of Architecture, Belgrade, 11000, Serbia, g) Polytechnic University of Madrid, Madrid, 28040, Spain, h) University of Information Science and Technology "St. Paul the Apostle", Ohrid, 6000, Macedonia, i) Fraunhofer Institute for Solar Energy Systems ISE, Freiburg, 79110, Germany, j) University of Agder, Grimstad, 4879, Norway, k) PV Performance Labs, Freiburg, 79110, Germany

***Abstract* - COST Action entitled PEARL PV aims at analyzing data of monitored PV systems installed all over Europe to quantitatively evaluate the long-term performance and reliability of these PV systems. For this purpose, a data bank is being implemented that can contain vast amounts of data, which will enable systematic performance analyses in combination with simulations. This paper presents the development process of this data bank.**

***Index Terms* — PV systems, Data monitoring, Data analysis, Performance, Reliability.**

I. Introduction

The objectives and background of COST (Action PEARL PV have been introduced in detail in [1]. This research network aims to increase performance and lower costs of electricity produced by photovoltaic (PV) solar electricity systems in Europe via (i) obtaining higher energy yields, (ii) achieving longer operational life time (beyond the 20 years usually guaranteed by manufacturers) and (iii) lowering the perceived investment risk in PV projects. These objectives will be achieved by a cooperative European COST Action partnership, collating and analyzing a very large aggregated set of PV system operational performance data, with a focus on understanding defects and failures of PV systems installed across Europe. This is organized in the context of integration of PV systems and components into grids and the built environment, and the impact of regional climate characteristics on the generation of PV energy. For this purpose, five Working Groups have been set up that will conduct research using a shared data bank and shared simulation tools and models to analyze and compare data that are collected in this data bank, see Figure 1. The core focus of this paper is the central facility of this Action: the data bank. The data bank has been in preparation since October 2018; subsequently its implementation has started in January 2019. In this paper, in Section 2 the considerations regarding the type of data bank and its system architecture will be presented and in Section 3 the expected research activities with various data sets will be presented.

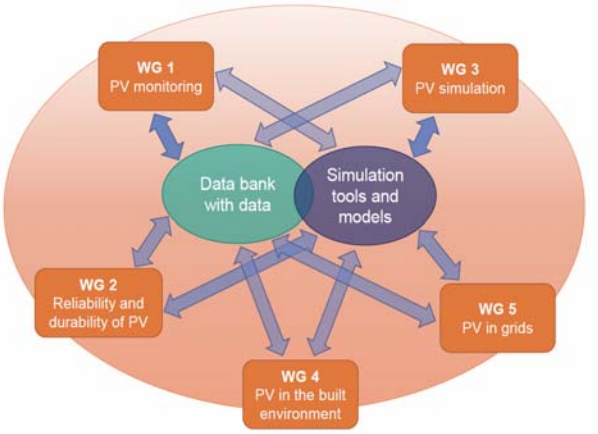

Fig. 1. The 5 Working Groups of COST Action PEARL PV in relation to a shared data bank and simulation tools.

II. Selection of the Data Bank Platform

This section will provide a short overview of different types of data bank platforms and structures and a logical explanation for the selection of CKAN (Comprehensive Knowledge Archive Network) [2] as the most appropriate choice for PEARL PV's data bank.

Since the beginning of the digital age, many developers have been developing platforms and structures in order to store data safely and reliably. Time-series data typically are challenging due to the vast quantity of data that has to be dealt

with. The database solution for the purposes of the project needs to be able to handle system performance data and its metadata and be able to link the metadata with the very large and dynamic time-series data. A big challenge is searching and analyzing the data, which could be done online at the platform or offline, after having downloaded the data, e.g., via scripts in any software environment. The system should also run as autonomously as possible. To find a good solution to store and query time-series data, we first considered several available approaches to the databases.

In relational SQL (structural query language)-based databases (e.g. MySQL, Oracle DB, SQL Server, PostgreSQL), data, once inputted, are abstracted from the users, so they have no direct access to it and the only possible access to the data is through the application software or queries. SQL based software stores data in tables. Normal relational databases perform well in storing data but perform poorly with queries when it comes to big data. NoSQL databases combine database elements with object-oriented programming languages. Instead of storing the data in tables, object-oriented databases store complex data objects in the database. There are plenty of NoSQL databases that are used for different purposes, which can be divided into four common types: key-value, column-oriented, document-oriented, and graph databases. We considered a few NoSQL databases approaches for manipulation of time-series data: MongoDB (document-oriented database), Apache Cassandra (key-value database), Scylla (compatible with Apache Cassandra), and Apache Hbase Hadoop database (column-oriented database). All these approaches are also doing well with storing of data, but do poorly with queries.

To overcome the limitations of the NoSQL databases related to time-series data, the concept of time-series databases is adopted. These databases are used to handle and manipulate time-series data. We considered InfluxDB, as it is one of the most popular time-series databases. InfluxDB provides an SQL-like language written for time-series data. One possible considered solution for our system was a combination of an InfluxDB database to store the data and a MongoDB database to store the metadata. HDF5 format is a self-describing data format capable to manage data collections of different sizes and complexity. It can store different types of data and their metadata together into one package. HDF5 files can be easily read into Python, R, MATLAB or any other data analysis language. Mondas [3] is an example of the software for the management, analysis, and visualization of time-series data, based on the HDF5 storage format. Another possibility is to use RedShift, which is part of the Amazon Web Services, as a flexible and cost-effective solution. The GUI (Graphical user interface) would then have to be developed in order to access the time-series data and the metadata. Other functioning PV data repositories are the DuraMAT Data Hub [4], which is built as a CKAN application, and PVOutput [5].

The chosen database structure and platform must meet the application requirements but also has to adapt to the expertise of the developers and database users (predominantly researchers). A major decision factor from the user point of view is that data do not have to be of a uniform format when uploaded, which helps to lower the barrier for contributors and will keep the door open for useful datasets that were not foreseen. Furthermore, the expected data size will be approximately 4TB to be used by 200-500 users from more than 30 European countries. Data uploads must be accompanied with meta-data and the system has to support this. The database requires search functionality, and one must be able to conduct searches based on metadata.

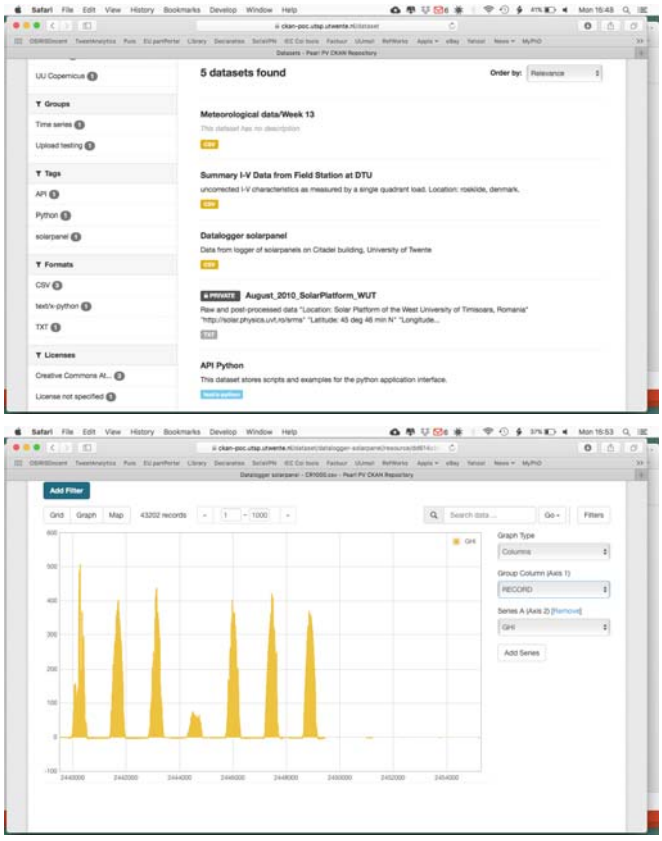

Fig. 2. Screenshots of datasets available on the proof of concept database (top), and an example graph that can be made online (bottom).

Preliminary research and the consideration of the requirements led to two possible candidate systems: Dataverse [7] and CKAN [2]. Both are used in systems which collect large scale solar data and are considered capable of handling ‘big data’ and associated meta-data. Based on its promising architecture, and implementation for the DuraMAT Data Hub [4], the CKAN system [2] was selected as a data bank for a proof of concept (PoC) implementation for this project. In theory, CKAN does not have any limits on the amount of data that it can store. However, it might experience operational

issues if its underlying systems, like the PostgreSQL database or storage system, hit their physical limits.

The implementation of the proof of concept has been done in small successive iterations, based on an Agile development process. A small team initially tested the functionality of the CKAN proof of concept implementation by uploading and downloading small data sets. Note that non-disclosure agreements (NDAs) have been signed prior to allowing access to the data base. After resolving minor issues, some 20 participants active in Working Group 1 of PEARL PV have been invited to perform more tests, again after signing NDAs. An example of the proof of concept data and graphing capabilities is shown in Figure 2. By mid-May 2019, the tests were found to be successful, and a full implementation was started. A database server (4 CPUs, 16 GB memory, max 4 TB diskspace) was installed successfully, on which CKAN is running. A scheme with relations between the data and metadata is shown in Figure 3.

III. Metadata

A questionnaire has been developed and distributed to all PEARL PV participants in order to support and make choices in the type and amount of metadata that will be collected. Analysis of the results has led to a distinction between required data and desired ('nice to have').

Foremost, quality control of the data is required. This means that accuracy of instruments and/or sensors must be specified. Also, gaps in data should be identified. Most respondents prefer 'csv' file format of the data, with missing data points identified as "NaN". Time format should be YYYY:MM:DD HH:MM:SS, in UTC.

Regarding metadata for PV systems and components, the following data is required: site name and GPS coordinates (latitude, longitude), type of PV installation (fixed, tracking, BAPV/BIPV, roof/façade, etc), for fixed system tilt angle (0-90º), and azimuth angle (0-360º, North being 0º), for tracking system 1-axis (E-W), 1-axis (N-S), or 2-axis, type of PV module technology, PV module characteristics (I_{sc}, V_{oc}, P_{mpp}, I_{mpp}, V_{mpp}, temperature coefficients), string design (number of modules, number of strings, connections to each inverter), shading (no shading), type of inverter technology (central, string, micro, transformerless, etc), inverter specifications (AC/DC power, frequency, rated efficiency, number of phases, reactive power, number of independent MPP's, total number of inverters, communication protocols). When applicable, the following BOS-components should be specified: battery-system and own developed hardware/software. Information about other parts of the BOS is desired, not required.

For the meteorological instrumentation, details and accuracy of pyranometer, air temperature sensor, PV module temperature sensor are required. Wind and other meteorological sensor information is desired, not required.

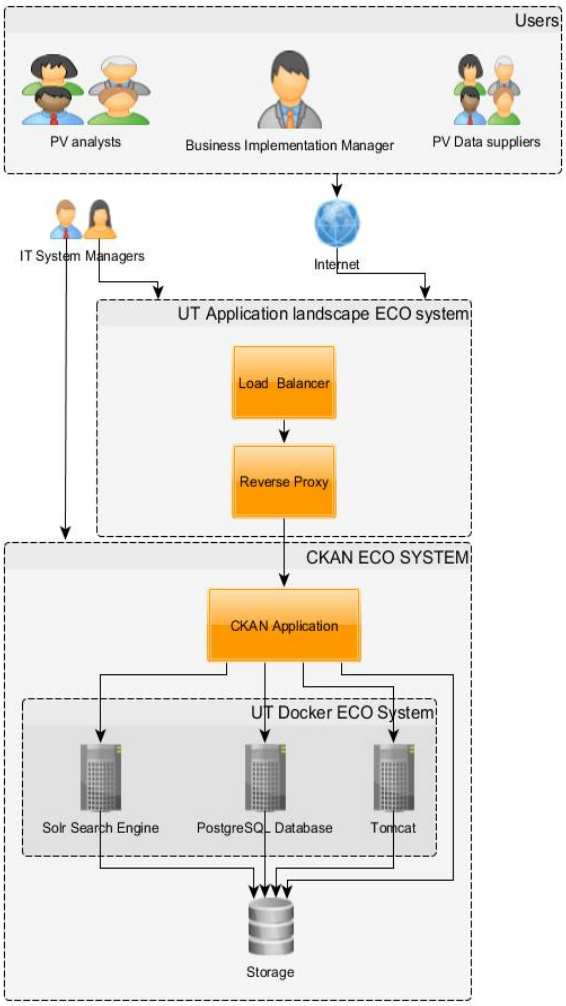

Fig. 3. Data bank structure using a CKAN Eco System.

For other instrumentation, the details and accuracy of power meters (DC and AC) and system status information (system downtime, sensor failure, cleaning events) is required. Metadata for IR and EL imaging and specific purpose irradiance sensors is desired, not required.

For monitoring of meteorological parameters, the required data is: global horizontal irradiance, global plane-of-array irradiance, air temperature. The required sample rate is 1 sec and 1 min, with a record rate of 1 min. Other parameters (diffuse/direct/spectral irradiance, albedo, air pressure, wind speed/direction, humidity, precipitation) are desired, not required.

For monitoring of yield and durability from PV modules/strings, the required data are: ambient temperature, module temperature, plane-of-array irradiance, module/string I_{mpp}, V_{mpp}, and P_{mpp}, sampling time. In terms of datalogging, a minimum time series of 1 month and 1 year should be available, with a sample rate of 1 min and 1 hr, and a record

rate of 1 min and 1 hr. Sample and record rates less than 1 sec are not required. For monitoring of degradation of PV modules/strings, the required data is: IV-curves recorded several times a day and once a week, cleaning frequency of PV modules and irradiance sensors (soiling), visual inspection and leakage current detection. Other failure or degradation indicators (corrosion, IR/EL/UV-FL images, or combined methods) are desired, not required.

For PV simulation related to the work of Working Group 3, the required data is: weather and irradiance input files (synthetic/TMY files, ground recorded data), and datasheets for the components modelled. The components required to be modelled are the inverter and the PV system. Models of PV modules/strings and BOS components (except inverter) are desired, not required. The simulation resolution should be 1 min and 1 hr. Other aspects, such as modelling the performance, external influences, thermal properties, building integration, and combined models for electrical/thermal/building properties, are desired, not required.

For monitoring of PV in the built environment, linked to the work in Working Group 4, the required data is: technical integration aspects (specification of BIPV module and fastening system, material choices, weatherproofing, installation procedures, operation and maintenance, electrical safety), and thermal considerations (ventilation of modules, temperature of modules, thermal properties, indoor temperatures). Information about architectural integration is desired, not required. Required data for other aspects are: Hybrid energy system (type of hybrid system, energy share and power share profile of the hybrid system), environmental footprint (energy payback time), and economic considerations (system costs, economic models, economic value of distributed energy production). Life cycle analysis and the added value of building integration is desired, not required.

For monitoring of PV in grids related to the work in Working Group 5, the required data is: inverter output power factor, current and voltage on each phase, active power of load, active power to/from grid, net active power and net energy (generation-load matching). For applications with energy storage, required data is: Operating voltage of energy storage, current to/from energy storage, and active power to/from energy storage. The required sample rate is 10 msec (sinewave), 1 sec, 1 min and 15 min, with similar record rates (10 msec, 5 sec, 1 min and 15 min). Information about reactive power flows, grid frequency and voltage stability, harmonics, and specific load characteristics is desired, not required.

IV. Use of PEARL PV's data bank for PV Research

Each Working Group (WG) of COST Action PEARL PV will use the data bank for joint PV research which will be partially based on the recently developed "Workplan 2018-2021" [8]. Below the research activities are presented for all working groups.

A. WG1 Research activities: PV Monitoring

The overall objective of Working Group 1 is to investigate long-term PV performance and reliability. The basis for this was setting requirements for essential data and nice-to-have data that should be entered in the data bank, which is described in the previous section. Next, data will be analyzed of the actual monitored long-term performance, defects and failures in PV systems installed all over Europe to quantitatively determine the absolute influences of components rated performance, key design of systems including BIPV, residential, field-based and floating systems, installation, operation, maintenance practice, geographic location and weather/climate factors on the performance, performance degradation over time and failure modes of these PV systems. Guidelines for analyses have been recently reviewed, and will be used as basis for analysis [9]. Close collaboration with IEA-PVPS-Task 13 is foreseen and guaranteed as many PEARL-PV participants are also active within IEA-PVPS-Task 13 [10].

B. WG2 Research activities: Reliability and Durability of PV

The main objectives of Working Group 2 are to define reliability and durability metrics for PV modules, components and systems, to identify relevant data to be collected to measure reliability and durability, and to share knowledge via workshops, seminars and joint publications originating from WG2 with a wider community of PV experts and other experts working for insurers, investors and banks. As today, on the one hand, different views of PV stakeholders (consumers, investors, manufacturers, researchers, utilities) exist describing reliability and durability, while on the other different metrics for different PV technologies (c-Si, thin film, organic....) are considered, a common metric would be beneficial. A white paper reviewing all this is planned. Further, identification of relevant data will include the description of mass PV data analysis methods - output power over time and multi-faceted analysis to gauge output decrease, the identification of issue causing decrease in power output (e.g. shading, physical degradation...) and a correlation of failure modes with climatic conditions.

C. WG3 Research activities: PV simulation

Working Group 3 considers the use of modelling tools to simulate the performance of photovoltaic devices and systems. This covers both the prediction and the assessment of performance and complements the activities of the other Working Groups in the PEARL PV project, especially those considering specific PV applications. The objectives of WG3 are first to classify PV simulation models by content ranging from (i) fundamental solar cell research, (ii) PV irradiance modelling including forecasting and cloud formation, (iii) PV

systems (grid-connected, stand alone and hybrid), (iv) PV in the built environment and (v) PV grid interactions, and second to compare these various models in various cases linked to the work in other WGs.

D. WG4 Research activities: PV in the built environment

One of the main targets for WG4 is to identify that part of collected big data that could be used in helping architectural and urban designers in general, to understand and adopt BIPV technology, and based on the empirical experiences, gain confidence in both the design potential and financial feasibility that could clearly be communicated to the clients in initial stages of the design process. WG4 will focus on the data coming from various built contexts, i.e. urban and rural environments from which information on urban morphology and discrete urban geometry can also be obtained. Integration in a 2D GIS, 3D GIS and BIM systems will be examined and implemented. The data of particular interest are derived from BIPV (Building Integrated PV) systems, and are primarily distinguished by the type of building and the position of PV facilities (rooftops, facades, shades, window glazing, etc.).

In the domain of BIPV data, possible case studies are planned, especially considering the so-called landmark objects, i.e. differing from its urban context (by size, geometry, social importance, building technology, etc.).

E. WG5 Research activities: PV in Grids

WG5 has started to analyze a subset of data that contains the energy production data and the metadata of about 20,000 PV systems in Europe (mainly France and Belgium). For about 6,000 of these PV systems, the energy production data has been recorded from 2011 to 2018 with a 10-min time resolution. Several tools will be developed by WG5, in connection with WG1, over the course of the next years. Peer-to-peer prosumer cooperation, e.g. using blockchain technology are increasingly appearing as a viable option for the future development of PV generation. Therefore, several tools will be developed: (1) spatio-temporal forecasting using data from distributed PV systems, (2) assessment of the PV power mitigation potential from the geographic dispersion of PV systems, (3) assessment of PV power fluctuations for PV system fleets, including the correlation between neighboring installations, and (4) the evaluation of the Power Quality indicators at the connection of PV systems to the grid. Studies will be conducted on the relationship between the PV production and the local consumption, the possible use of batteries, or the economic viability of alternative options. Also, a fault detection toolbox will be developed to improve the energy yield of grid-connected PV systems and reduce their power instability.

V. Conclusion

With the realization of a database structure and platform presented in the paper, the many research questions that have been formulated in the various working groups can be answered in the coming years. Of course, substantial amounts of data are needed to be supplied by the PEARL PV participants, and others as well. Therefore, in the upcoming months all participants will be encouraged to share data. After signing an NDA, all PEARL PV participants will have access to all data sets in the database, to explore and analyze as they wish. Of course, proper acknowledgement to the source of the data is required in publications based on the data. Having available data from various sources, we expect that this will lead to increased collaborations and the development of standardized, and open, analysis tools.

Acknowledgements

We would like to thank all 35 member countries and nearly 200 participants of PEARL PV for their enthusiasm and efforts. In particular the whole LISA team of the ICT department of University of Twente, specifically Guido van de Zweerde, are greatly thanked for the development of the data bank. This work is based upon work from COST Action PEARL-PV, supported by COST (European Cooperation in Science and Technology), see www.cost.eu.

References

[1] A. Reinders, D. Moser, W. van Sark, G. Oreski, N. Pearsall, A. Scognamiglio, J. Leloux "Introducing 'PEARL-PV': Performance and Reliability of Photovoltaic Systems: Evaluations of Large-Scale Monitoring Data" in Proceedings 7th WCPEC, 2018, pp. 762-766.

[2] CKAN Association. (n.d.-a). CKAN code architecture — CKAN 2.8.2 documentation. Retrieved May 30, 2019, from https://docs.ckan.org/

[3] Mondas, http://www.mondas-gmbh.de/, last accessed May 30, 2019

[4] DuraMAT Data Standards and Guidelines. (2013, September 25). Retrieved May 30, 2019, from https://datahub.duramat.org/dataset/

[5] PVOutput, https://pvoutput.org/, last accessed May 30, 2019.

[6] Dataverse, https://en.wikipedia.org/wiki/Dataverse, last accessed May 30, 2019.

[8] PEARL PV Workplan 2018-2021, Version 2: 18 November 2018, https://www.pearlpv-cost.eu/wp-content/uploads/2018/11/CA16235-Work-Plan-2018_2021-_18112018.pdf, last accessed May 30, 2019.

[9] B.R. Paudyal, A.G. Imenes, T.O. Saetre, "Review of guidelines for PV systems performance and degradations monitoring", Proceedings 35th EUPVSEC, 2018, pp. 2037-2050.

[10] IEA PVPS Task 13, http://www.iea-pvps.org/index.php?id=57, last accessed May 30, 2019.