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], Nicola Pearsall[e], Mirjana Devetakovic[f], Jonathan Leloux[g], Dijana Capeska Bogatinoska[h], Anton Driesse[i]

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) V Performance Labs, Freiburg, 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 of these PV systems. For this purpose a data bank is being implemented which can contain big data and which will enable systematic analyses in combination with simulations. This paper presents the development process of this data bank as well as the first analyses results of various data sets.**

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 defect and failure of PV systems installed across Europe, in the context of integration of PVs facilities into grids and the built environment, and the impact of regional climate characteristics on the generation of PV energy. For this purpose, 5 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 these 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.

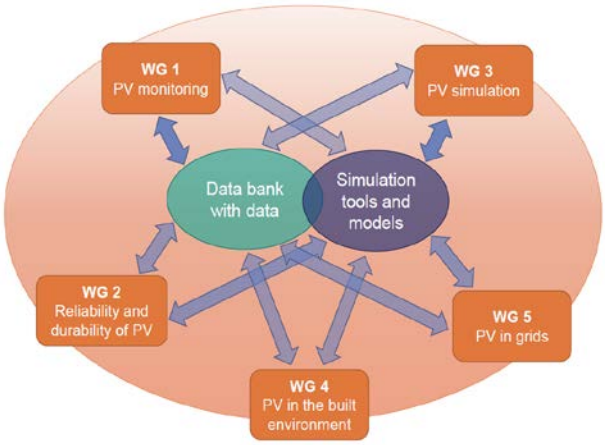

Fig. 1. The 5 Working Groups of COST Action PEARL PV in relation to a shared data bank and simulation tools.

II. Selection of the Data Bank Structure

This section will provide a short overview of different types of data bank structures and a logical explanation for the selection of CKAN as the most appropriate choice for PEARL PV's data bank.

Since the beginning of the digital age, many developers have been struggling to store data safely and reliably. Time-series data present its own challenges due to the vast quantity of data that has to be dealt with. This data requires significantly more flexibility, agility, and scalability than it is the case in standard PV data analyses. The database solution for the purposes of the project needs to be able to handle the metadata and be able to link it with the very large and dynamic time-series data. A big challenge is searching and analyzing the data. 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 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 are doing well with storing the data but do 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 4 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 file can be easily read into Python, R, MATLAB or any other data analysis language. Mondas [2] is an example of the software for the management, analysis, and visualization of time-series data, based on HDF5 storage format. Another possibility is to use RedShift, a part of the Amazon Web Services, as a flexible and cost-effective solution. The GUI (Graphical user interface) would then have to be developed to access the time-series data and the metadata. Other functioning PV data repositories are the Duramat Data Hub which is built as a CKAN application [3] and PVOutput [4].

The chosen database must meet the application' requirements but also has to adapt to the expertise of the developers and database users (e.g. 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 keeps the doors 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, the meta—data model is known and the system has to support this. The database needs search functionality, and one must conduct searches based on metadata.

Preliminary research and the consideration of the requirements led to two possible candidate systems: Dataverse [5] and CKAN [6]. 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 [3], the CKAN system (CKAN Association, n.d.-b) was selected as a data bank for the 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 run into operational issues if its underlying systems, like the PostgreSQL database or storage system, hit their limits.

The implementation of the proof of concept will be done in small successive iterations, based on an Agile development process. After each iteration a milestone moment is planned, with a demo of the proof of concept and a discussion with all the stakeholders. The stakeholders will decide the direction of development of the proof of concept, which will also be used to steer the planning. If the proof of concept is successful, the development will continue with the full implementation of a prototype to be finished in April 2019. A scheme with relations between the data and metadata is shown in Figure 2.

III. 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" [7]. Below research activies are presented for three WGs.

A. WG1 Research activities: PV Monitoring

The overall objective of Working Group 1 is to investigate long-term PV performance. This will be achieved by first setting requirements for essential data and nice-to-have data that should be entered in the data bank. 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. Close collaboration with IEA-PVPS-Task 13 is foreseen and guaranteed as many PEARL-PV participants are also active in with IEA-PVPS-Task 13.

B. *WG4 Research activities: PV in the built environment*

One of the main targets for WG4 is to identify the part of collected big data that could be used in helping architectural and urban designers in general, to understand and adopt the BIPV technology, and based on the empirical experiences, gain a confidence in the both design potential and financial feasibility that could clearly be communicated to the clients in initial stages of 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 is to 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.).

C. *WG5 Research activities: PV in Grids*

As January 2019, the 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). On 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 during 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.

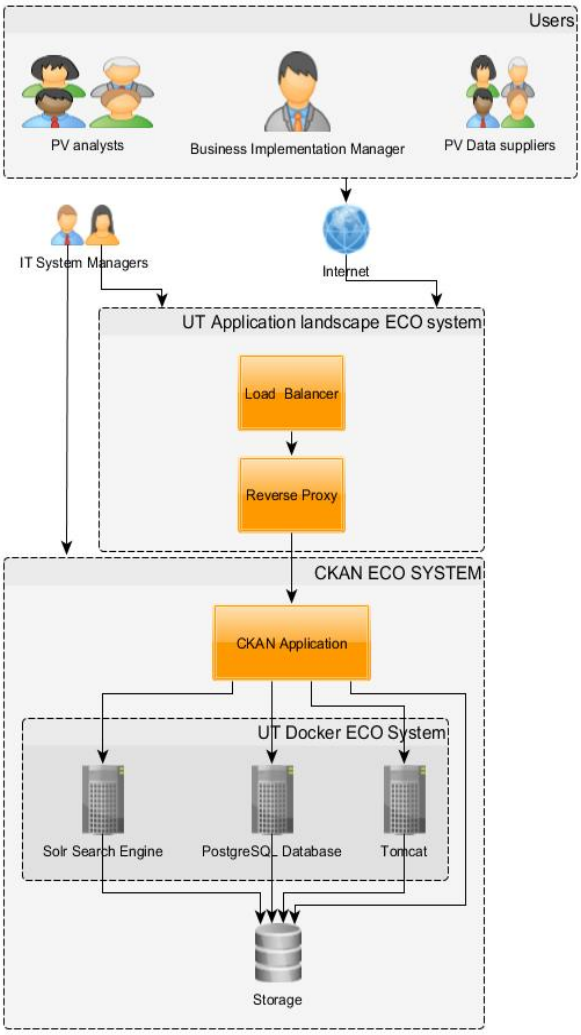

Fig. 2. Data bank structure using a CKAN Eco System.

ACKNOWLEDGEMENT

We would like to thank all 190 participants of PEARL PV for their enthusiasm and efforts. In particular the whole LISA team of the ITC department of UT are greatly thanked for the development of the data bank. This abstract 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 and J. Leloux "Introducing 'PEARL-PV': Performance and Reliability of Photovoltaic Systems: Evaluations of Large-Scale Monitoring Data" in 7th WCPEC, 2018.

[2] Mondas, http://www.mondas-gmbh.de/, 2019

[3] DuraMat Data Standards and Guidelines. (2013, September 25). Retrieved January 21, 2019, from https://datahub.duramat.org/dataset/

[4] PVOutput, https://pvoutput.org/, 2019.

[5] Dataverse, https://en.wikipedia.org/wiki/Dataverse, 2019.

[6] CKAN Association. (n.d.-a). CKAN code architecture — CKAN 2.7.3 documentation. Retrieved January 20, 2019, from https://docs.ckan.org/en/ckan-2.7.3/contributing/architecture.html

[7] 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