PACT Data Management Plan

Version 1.0

Robert White[1], Kirsten Perry[1], Bruce King[2]

[1] National Renewable Energy Laboratory

[2] Sandia National Laboratories

October 2021

SAND2021-13254 R

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

PACT Data Management Plan

Version 1.0

Robert White[1], Kirsten Perry[1], Bruce King[2]
[1] National Renewable Energy Laboratory
[2] Sandia National Laboratories

October 2021

CONTENTS

LIST OF FIGURES

EXECUTIVE SUMMARY

In order to meet the guidelines, set forth in the US Government Open Data Initiative and data management requirements set forth by DOE, we shall implement the following procedures to protect and curate all project or consortium data. Instrumentation data will be backed up locally on redundant hard drives, onto cloud-based repositories, or on local laboratory information management systems to protect raw recipe, operations, time-series, and characterization data.

Data products will be aggregated to an NREL-operated Data Hub for the purpose of access and release to consortium project members, customers, and possibly the public, once any proprietary or publication embargos have been rescinded. All data found to be of business confidentiality or intellectual property shall be maintained in closed projects, with access granted only to authorized members. Any data containing Personally Identifiable Information (PII) or with national security implications will not be stored or released in accordance with all laws and DOE regulations, orders, and policies. Data that is to be archived for the consortium shall include, but is not limited to, all data that has been contextualized and found useful to support research and reporting to customers and partners, and any data needed to validate and reproduce any published results.

Access to project level proprietary/private data on the Data Hub requires two-factor authorization and project access granted by the project owner or lead or system administrators. By default, all data uploaded to the Data Hub will be set as project private.

At the end of instrumentation and experimental lifetime or the shutdown of the consortium, all data products will be placed in a master archive to preserve it for future needs. After shutdown and in the event of suspended Data Hub operations, data products can be made available through direct request to the project lead or their designated representative.

1. INTRODUCTION

The Perovskite PV Accelerator for Commercial Technology (PACT) is an independent validation center for the evaluation of perovskite PV technologies and their bankability. The center is led by Sandia National Laboratories and the National Renewable Energy Laboratory (NREL) and includes as part of its team Los Alamos National Laboratory (LANL), CFV Labs, Black and Veatch (B&V), and the Electric Power Research Institute (EPRI). The goals of the center are to:

- Develop and improve indoor and outdoor performance characterization methods
- Develop and validate accelerated qualification testing for early failures (5-10 years),
- Research degradation and failure modes
- Validate outdoor performance
- Provide bankability services to US perovskite PV (PSC) industry

The importance of data and data management to the success and outcomes of the PACT center is paramount. This report describes how data will be managed and protected by PACT and identifies important data management principles that will guide our approach.

1.1. FAIR Principles

Whenever possible the consortium will seek to adhere to the FAIR principles for scientific data management and stewardship[1]. These principles promote the ideas of managing data so that it is fully contextualized, allowing the raw data to be easily used and reused. The concepts of the FAIR principle are:

- ***Findability****:* The data and associated metadata is persistent, richly described and can be queried through machine access (Application Programmable Interfaces or API) or through manual interaction.
- ***Accessibility****:* The data can be accessed easily and does not use proprietary access methods, even when allowing for authentication and authorization.
- ***Interoperability****:* Data is not stored in proprietary formats, and the data formats are widely accepted and easily read by a variety of programming interfaces. Common community vocabularies will be used to help describe any data and metadata.
- ***Reusabilit****y:* The metadata shall be clear and concise, following community standards and establishing provenance, that allows for interpretation and utilization in other methods beyond the original conceived scope.

[1] https://doi.org/10.1038/sdata.2016.18

2. DATA MANAGEMENT SYSTEM

2.1. Architecture

The PACT data ecosystem will integrate feeds from both manual web applications and a series of databases, previously deployed or under development, to house and archive data from a variety of instruments and testbeds important to PACT research. Some of the databases and associated file system archives are currently in operation and others are in the design and testing phase. The following diagram (See Fig 1.) shows the data flow from the instrumentation to the point of access by the PACT research team. A description of each of the data repositories follows.

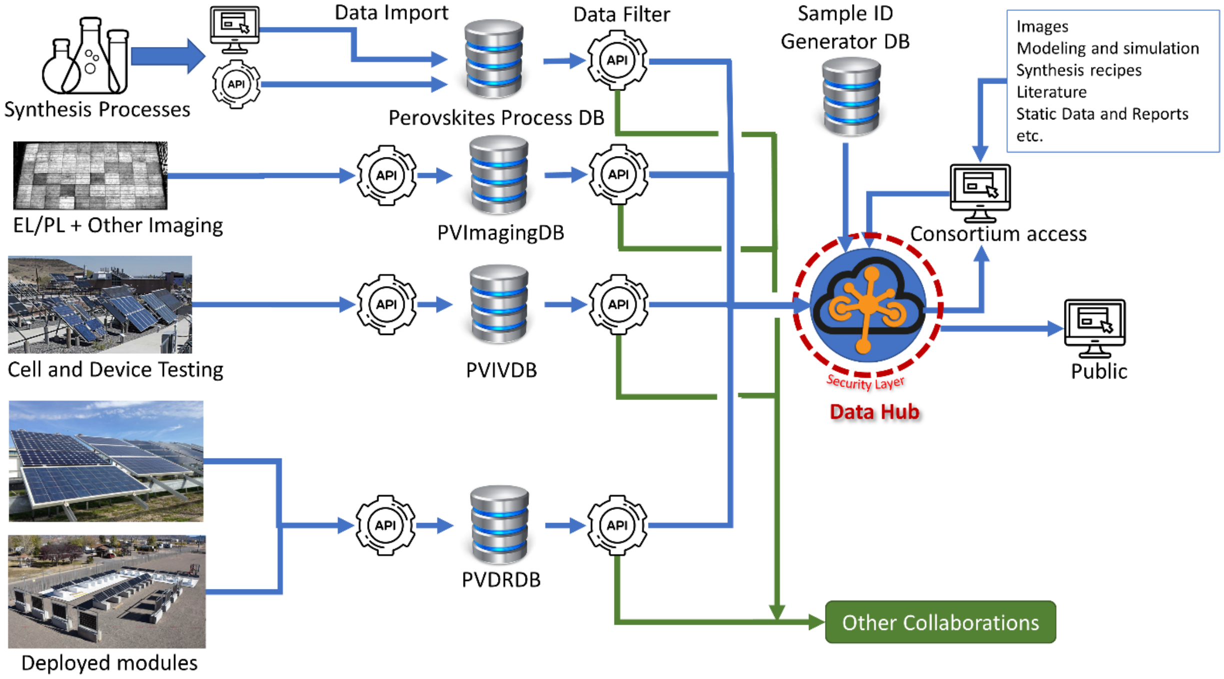

Figure 1. An example architecture to demonstrate the flow of data from the instruments to the data hub, and researchers and partners. Additional databases or sources from future instruments can be easily added to the data ecosystem.

- **Perovskites Process DB** – This database is used to gather information about the synthesis of perovskites cells and devices. The architecture and tables are based around the International Summit on Organic Photovoltaic Stability (ISOS) protocols[2] , or similar protocols. The data can be processed in through input of standard ISOS protocol Excel

[2] https://doi.org/10.1038/s41560-019-0529-5

file, or possibly through cloud-based documents (e.g., Google Docs/Forms). While much of the data from synthesis may not end up directly relevant to PACT, some of the results could be important to PACT. Access would be provided by requested uploads to the Data Hub on a case-by-case basis.

- **PVIVDB** - This AWS Redshift database stores time-series IV curve measurement data from field deployed cells and modules. This includes instrument setup and configuration, IV curves, as well as associated performance measurements (e.g., temperature, irradiance, etc.). The database is customized to support time-series and is scalable and agile in both data loading and queries. Data will be filtered and possibly anonymized before upload to the Data Hub.
- **PVImagingDB** - This database, under development, will act as an image repository. It will contain the links to archived images and details concerning the instrument and module setup during imaging. The database will include electroluminescence (EL) and photoluminescence (PL) images, as well as other possible types of images including scanning electron microscope (SEM) images and atomic-force microscope (AFM) scans. PACT-specific modules would be filtered from the broader PVImagingDB database for upload to the Data Hub.
- **PVDRDB** – PVDRDB is a large, scalable AWS Redshift time series database of system performance from fielded modules. This database has been in operation since 2019 and can be utilized to store module performance measurements with high time granularity. It has a suite of Python tools already constructed to facilitate automated data loading and querying. PACT-specific modules under test would be filtered from the broader PVDRDB datasets for upload to the Data Hub.
- **Sample Database** – This is a general, universally accessed database to house basic metadata for any sample submitted to or developed by PACT. A key element will be the generation of a standard sample ID based on information submitted to the database. This sample ID will be a key to linking data across all the other databases within the data ecosystem.
- **The Data Hub** – This is a software framework that consists of user interfaces, file system archives, and a high-level database to monitor transactions and to archive and disseminate information aggregated from the specific databases or through direct researcher contributions. Any data and metadata stored within has been filtered and possibly anonymized from sources within the data ecosystem to contain only PACT-specific datasets.

2.2. Data Acquisition Details

PACT data will be filtered through the main repositories from a variety of different instruments and sources. In many cases, these instruments are supporting both PACT and other projects or collaborations. In most cases, each of these instruments or groups of instruments will have their data harvested into databases supporting that instrument or area of research. We will provide methods for the researchers to load data to the Data Hub by either automatically acquiring data by scanning the instrument databases for relevant data sets, or by providing web applications or executable Graphical User Interfaces (GUI) for manual processing.

A tool set will be constructed to provide an Application Programming Interface (API) to perform an extraction translate and load (ETL) process on any data and metadata stored within instrument databases and the Data Hub. All data and metadata will be keyed on the relational tables through the standard sample IDs and timestamps. Data will be filtered from the databases by PACT-specific scripts utilizing these APIs, to enable scanning, processing, and possibly anonymizing the data before pushing the data up to the Data Hub.

At the most basic level, anonymization means removing the association of a manufacturer/customer, a particular technological process, or synthesis recipe from the shared data. We expect that some manufacturer/customers could have different requirements on what to secure, and we will make sure that we can meet or exceed those levels of security. As an example: A particular module would be associated to an ID number. Only by special authorization could a researcher gain information on who developed the sample module, how it was constructed, and of what materials.

Besides the main instrumentation data, there is additional ad hoc data that PACT could need or find useful. The Data Hub can provide single file or bulk file uploads of data to any general project, or any project associated with a sample. This could include data like modeling and simulation results, literature searches, system images, synthesis protocols, measurement benchmarks, reports, and more.

2.3. Sample Tracking

A separate database will be constructed and linked through the Data Hub that can store general sample metadata (customer, composition, request date, etc.) and generate standardized sample ID numbers, which will be used to track samples sent to PACT. These same ID numbers will provide a reference point (or primary key) in each of the instrument databases. This kind of architecture will allow for querying all procedures and data associated with any sample at a particular time, or over the sample lifetime.

2.4. Data Security

Data is protected within the data hub through layered architecture. The data hub utilizes a Cognito user pool that will need to be initialized upon stand up of the production version of the hub. This is a moderate level deployment, and therefore cannot utilize the existing Cognito user pools established for the Energy Material Network Data Hubs. Access to this data hub will require the user pool to provide a two-factor authorization to log in. The users will be able to utilize any current available off-the-shelf smart phone or computer-based applications (e.g., Authy, Google Authenticator) to receive the two-factor code.

Since all data within the Data Hub is private by default, this provides a "need-to-know" access level that can be controlled by a specific project lead or the system administrators. Once access for a logged-in user has been verified and established, any project data the researcher can access will be available through project or data pages. A project lead can remove or add a user to access the data, dynamically, at any time.

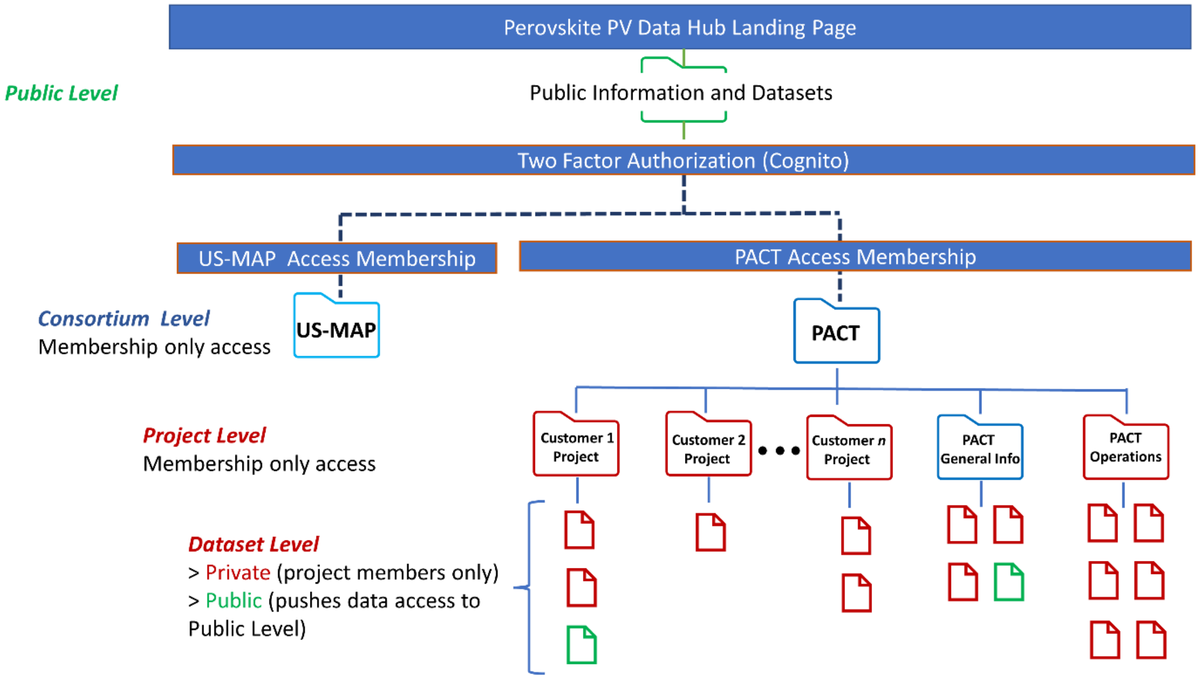

Figure 2. A diagram of the layered data security. Access to the landing page and any publicly released data sets are available to all people arriving on the data hub site. Authorization to PACT or US-MAP specific areas of the datahub requires a two-factor login. Access to consortium, project, or dataset level is based on the logged in researcher's memberships.

It is also important to note that each of the databases that feed the Data Hub will have their own access security, which is typically limited to a handful of researchers that work directly with the equipment and are authorized by the instrument owner.

2.5. Data Access

Data is available to any researcher within PACT, based on their "need-to-know" status. By logging in, the researcher will then have access to all data in projects in which they are a member. Typically, any partner with a business-sensitive research goal will be given a project to house their data, to obscure the data from any competing users. Any project-sensitive data will be stored within and accessible only to project members and the partner.

Security is assigned at the project or customer level. The members of any project/customer folder can see who currently has access to the datasets, but access permissions are only allowed to be assigned by the PI or customer lead, or their representative. The permissions are dynamic and they have the ability to assign and remove other PACT members access to the datasets at any time. Only data that has been anonymized and vetted would be available to the general PACT membership.

Researchers can use the basic website application to download any accessible datasets. Each user is issued a unique API key once they become a registered user. This will be used programmatically for authentication, to utilize the website via the API. We will provide a series of tools and tutorials for users to use their API keys, facilitating access to create, search, upload, and download datasets without requiring manual access the Data Hub site.

2.6. Digital Object Identifiers (DOI)

If data is to be released to the public, the researcher has the option, and will be encouraged to, have a digital object identifier (DOI) assigned to the public dataset. This will allow for citing the dataset within publications and preserve the data provenance. The Data Hub will utilize the OSTI systems for creating of the DOIs.

2.7. Data Standards

Any data products stored within the data hub need to utilize simple standardized formats. Preferably all data will be recorded in ASCII-encoded CSV formats. These formats are easily loaded and adapted for data analysis. When CSV format is not available, an ASCII-encoded text file is the next best solution.

In the case of proprietary formats (.docx, .xslx, etc.), the Data Hub can store these files, but their capability at being easily utilized through API access or simple queries is greatly reduced if not eliminated.

Files that are produced by instrumentation that are proprietary and require special programs to access are a last resort. In such cases, the researcher uploading the file must include documentation stating the program details to be able to access the file, including the version, name, and software developer or manufacturer.

Simple image formats can be stored and directly viewed in the Data Hub without download. These include files with the extensions .png, .jpg and .gif. Other formats (.TIF) can be stored but are not viewable in the Data Hub. Possible file conversions can be made during upload to add viewable image formats along with the original complex, high resolution formats.

2.8. Standard Taxonomy

PACT will seek to establish a set of standard terminology that will be used to describe as many aspects of the research data and metadata as possible. This will include ways to label measurement types, procedures, and processes. This will allow for easily parse-able text sets to facilitate many automated processes and will also reduce confusion when researchers are discussing data and topics. More details can be found in the working glossary document found in Appendix A.

2.9. Researcher Requirements

While much of the Data Management plan covers many instances that can be maintained through engineering controls, every researcher in PACT will still need to manage the data associated with their work, to ensure proper database ingest and storage.

- Each researcher in PACT will be responsible for assuring they are pushing data to the Data Hub, for access by colleagues, partners, and customers as often as possible or needed.
- The researcher needs to be proactive in assuring that metadata that is required or optional for any measurement or analysis is properly addressed and added to any uploads to the Data Hub.
- The researcher needs to assure that their data is properly labeled, and column names match a standard glossary.
- The researcher needs to acquire and use a proper sample ID to tag any metadata file, data file, or documents related to a sample.

2.10. Project Termination

In the event of the termination of the PACT research consortium, all public data and any DOIs will be maintained through the existing Data Hub, but the Data Hub will be static and not accepting any additional submissions. All private datasets will be extracted and archived within the NREL Computational Sciences Data Center, and any needed data can be acquired through request from the project leads.

APPENDIX A. GLOSSARY AND ACRONYMS

AFM – Atomic Force Microscopy

API – Application Programmable Interface

ASCII – American Standard Code for Information Interchange

AWS – Amazon Web Services

Cognito – Amazon secure user authentication service

CSV – Comma Separated Variables

DOI – Digital Object Identifier

EL - Electroluminescence

EMN – Energy Materials Network

ETL – Extraction, Translate and Load

FAIR – Findability, Accessibility, Interoperability, Reusability

GUI – Graphical User Interface

ISOS - International Summit on Organic Photovoltaic Stability

IV – Current-Voltage

OSTI – U.S. Department of Energy Office of Scientific and Technical Information

PACT - Perovskite PV Accelerator for Commercial Technology

PL - Photoluminescence

Project – A grouping of specific activities or experimental thrusts. All data generated by a project is stored within that project folder on the data hub. Access to any protected project data is authorized by membership in the project.

Project lead – The owner or director of a project, as specified on the data hub. This individual will be responsible in approving access to any secured data within the project and approves of any data to be released to the public.

PSC – Perovskite Solar Cell

SEM – Scanning Electron Microscope

US-MAP – U.S. Manufacturing of Advanced Perovskites

PACT

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

U.S. DEPARTMENT OF ENERGY Sandia National Laboratories NREL NATIONAL RENEWABLE ENERGY LABORATORY Los Alamos NATIONAL LABORATORY BLACK & VEATCH EPRI ELECTRIC POWER RESEARCH INSTITUTE CFV Labs