

The Challenges of Curating a Research Data Set From the World's Most Complex Machine

Eric Andersen and Jeff Banning, Pacific Northwest National Laboratory

The US Department of Energy's Office of Electricity has supported research for big data analytics of phasor measurement unit (PMU) data as a step toward the development of real-time early warning tools, operator decision support tools, and potential PMU-based automated controls for the world's most complex machine: the U.S. bulk electric system. The goal of DOE's Funding Opportunity Announcement (FOA) 1861 was to explore the use of **artificial intelligence (AI)** tools on **time-synchronized telemetry data (synchrophasors)**, to confirm and improve existing knowledge, and to discover new insights and tools for better electric grid operation and management. This FOA offered pre-packaged datasets to the FOA awardees to develop AI tools and capabilities. Pacific Northwest National Laboratory (PNNL) was responsible for acquiring the electric utility data to be provided to awardees of the FOA call and to provide support with the execution of the projects and their outcomes.



Identify Utilities

Establish NDA's

Acquire Data & Event Logs

Load Raw Data onto Cloud

Clean Up and Anonymize Data

Data Delivered to Awardees for Use

Data Request:

- Geographically dispersed utilities from each of the 3 U.S.-based electric grid interconnections
- Two years of data

Challenges:

- NDAs with Data Providers
 - Legal negotiations
 - Providing assurance of data protection
 - Providing details of the anonymization process before we had access to the data

Potential Remedies:

- Get a better understanding of utility data sharing risks and mitigation strategies
- Improve the value proposition for utility participation

Challenges:

- Pulling archive data was a heavy lift for some Data Providers
- Duration and age of data requested resulted in some providers having to obtain their data from third party archives
- Exported data came from a wide variety of archive tools (commercial, open-source, and custom) and were in different formats
- Event logs were all unique and lacked common taxonomy, and appear to be created manually

Potential Remedies:

- Move toward standardizing the data retention and archive processes
- Automate the creation of event logs and improve their consistency

Challenges:

- Inconsistency of
 - data between providers (e.g., different sequences and phases)
 - Extracted data formats
 - UTC timestamp formats
 - significant digits
- Extensive manipulation was required to aggregate data
- Data quality varied by provider
- Duplicate data
- Anonymization required to remove topology metadata

Potential Remedies:

- Improve
 - consistency across utilities
 - archive processes
 - what data is archived
 - UTC formats and timestamps
 - Data quality

Path Forward:

- There have been numerous requests for access to the anonymized data set
- NDAs with the Data Providers are being modified to enable additional use of the anonymized data set
- Additional synchrophasor data is being added, where possible
- If and when we are able to provide the anonymized data set for others to use, the data users will have to establish an NDA with PNNL

Summary of Utility Synchrophasor Data Contributed

Interconnection	Dataset Range	Number of Data Providers (Utilities)	Total PMUs	Raw Data Size Received (TB)
Eastern	2016-01-01 – 2017-12-31	5	250	38.6
ERCOT	2018-07-21 – 2019-08-24	5	221	10.6
Western	2016-01-01 – 2017-12-31	3	43	19.0
Total		13	514	68.2