Local Energy Oxfordshire

March 2022 | Version 1.2

Project LEO Data Cleaning and Processing

Masaō Ashtine

University of Oxford
Oxford Brookes University
low carbon hub
Oxford City Council
Oxfordshire County Council
Scottish & Southern Electricity Networks
piclo
EDF ENERGY
NUVVE
Origami Power Over Energy
Industrial Strategy
UK Research and Innovation

Report Title:	Project LEO Data Cleaning and Processing
Author(s):	Masaō Ashtine
Organisation(s):	University of Oxford

Version:	1.2	Date:	10/03/2022
Workpack*:	WP4	Deliverable:	D4.23
Reviewed by:	Inga Doherty		
Date:	25/02/2022		
Signed off by:	David Wallom		
Date:	28/03/2022		

Can be shared (Y/N):	Internally	Y	Publicly	Y

Context

The UK Government has legislated to reduce its carbon emissions to net zero by 2050. Meeting this target will require significant decarbonisation and an increased demand upon the electricity network. Traditionally an increase in demand on the network would require network reinforcement. However, technology and the ability to balance demand on the system at different periods provides opportunities for new markets to be created, and new demand to be accommodated through a smarter, secure and more flexible network.

The future energy market offers the opportunity to create a decentralised energy system, supporting local renewable energy sources, and new markets that everyone can benefit from through providing flexibility services. To accommodate this change, Distribution Network Operators (DNOs) are changing to become Distribution System Operators (DSOs).

Project Local Energy Oxfordshire (LEO) is an important step in understanding how new markets can work and improving customer engagement. Project LEO is part funded via the Industrial Strategy Challenge Fund (ISCF) who set up a fund in 2018 of £102.5m for UK industry and research to develop systems that can support the global move to renewable energy called: Prospering From the Energy Revolution (PFER).

Project LEO is one of the most ambitious, wide-ranging, innovative, and holistic smart grid trials ever conducted in the UK. LEO will improve our understanding of how opportunities can be maximised and unlocked from the transition to a smarter, flexible electricity system and how households, businesses and communities can realise the benefits. The increase in small-scale renewables and low-carbon technologies is creating opportunities for consumers to generate and sell electricity, store electricity using batteries, and even for electric vehicles (EVs) to alleviate demand on the electricity system. To ensure the benefits of this are realised, Distribution Network Operators (DNO) like Scottish and Southern Electricity Networks (SSEN) are becoming Distribution System Operators (DSO).

Project LEO seeks to create the conditions that replicate the electricity system of the future to better understand these relationships and grow an evidence base that can inform how we manage the transition to a smarter electricity system. It will inform how DSOs function in the future, show how markets can be unlocked and supported, create new investment models for community engagement, and support the development of a skilled community positioned to thrive and benefit from a smarter, responsive and flexible electricity network.

Project LEO brings together an exceptional group of stakeholders as Partners to deliver a common goal of creating a sustainable local energy system. This partnership represents the entire energy value chain in a compact and focused consortium and is further enhanced through global leading energy systems research brought by the University of Oxford and Oxford Brookes University consolidating multiple data sources and analysis tools to deliver a model for future local energy system mapping across all energy vectors.

Table of Contents

1. Executive Summary

This report highlights work that has been done on Project LEO's tools for data cleaning and quality control since the last version of this report was published in March 2021[1]. This report does not outline the detailed steps used to clean and improve data in LEO, but the main processes and tools implemented within LEO, including links to other useful documentation. This update briefly touches upon the goals of these data cleaning and data quality tools, and how they open the access to data analysis to both internal and external stakeholders, including the challenges in data cleaning for effective data management[2]. The following sections also summarise the migration of data tools from more inaccessible formats such as Python scripts, to more accessible online dashboards that strip away programming elements, allowing users a more friendly and guided experience. Much of these improvements employ the use of Dash capabilities where all supporting documentation and scripts will be made publicly available (where appropriate), facilitating easier adoption by *fast-followers* in other local energy systems to improve data management.

2. The Importance of Cleaning

Data cleaning involves the systematic processing and filtering of data (largely in tabular/relational format) to ensure maximum data quality for further processing and analysis. Data cleaning, when automated in later stages, frees up a lot of human and computational resources within projects that handle 'big data' or sizeable datasets. The effective pre-processing of submitted data takes only seconds (barring the development of the algorithms driving the cleaning) to perform, saving data managers and subsequent users along the data chain many hours of tedious work to correct erroneous data, reformat datasets, or improve the interoperability of differing dataset types. Project LEO's diverse ecosystem of MVSs (Minimum Viable System) and partners leads to equally diverse datasets, methods and outcomes which modern and innovative data cleaning methods will address in keeping with the Data Standards and Protocols document[3].

LEO is a project that intends to create impact beyond its research and operation boundaries and all our tools have been developed with open-source software, access, and use in mind. Though tools are aimed towards LEO partners, with specific needs being met, the design of our cleaning tools can be implemented in a wide range of projects and the code will be made available through our data repositories at the end of the project.

[1] https://project-leo.co.uk/reports/data-cleaning-and-processing-march-2021/

[3] https://project-leo.co.uk/reports/standards-and-protocols-report/

3. The Challenges in Data Cleaning

Data cleaning, particularly in energy systems involving a diverse set of actors and assets, presents many challenges. First and foremost are the differing needs of the intended users who will have various reasons for performing data cleaning. For instance, within LEO, we have seen cases where missing time periods in metered data is of more concern to one asset operator than timestamp mismatch issues from two meters and their subsequent synchronisation that may be more important for cleaning for another asset owner. The handling of different data resolutions will also affect what methods are applied to clean the data with statistical confidence. For example, where cleaning is concerned, a dataset containing 1-second data for a 2-hour flex service window is very different from a 2-day dataset containing hourly energy data. Being able to produce a tool that can handle all of these (and more) challenges is resource heavy, but significant effort has been placed into the development of functionality that incorporates a wide range of dataset formats to meet needs as best as possible; LEO's tools must be viewed in this manner and with an understanding that they are starting points for data pre-processing.

4. Who controls cleaning?

Our work on data cleaning has led to an initial discussion around which party in a flexibility market is responsible for performing data cleaning on asset and substation data. Substation data provide a clearer answer, as DNOs are controllers of these data points and are thus held responsible for providing accurate and carefully (transparently) cleaned data to address any data gaps. However, where grid-edge assets and metering enter the debate, further examination of data cleaning responsibility is needed. SSEN for instance, may want to absorb and perform data cleaning within the Neutral Market Facilitator (NMF), Whole System Coordinator (WSC), and baselining protocols to standardise cleaning across all ingested and output datasets.

Jade
Hydro operator with high-resolution, high-quality data across various parameters

Akeem
Battery operator with high-resolution data with limited metering

An example to put the above in context is as follows: Akeem and Jade are both engineers in charge of the scheduling, dispatch, and monitoring of their respective assets; Jade a hydro operator and Akeem a small-scale battery operator. The hydro can provide historically long-term and high-resolution datasets on a variety of parameters around the asset's operations and performance. Greater data oversight and control will afford Jade more confidence in data cleaning, particularly as long-term historical data enable better forecasting and gap-filling of periods of missing data (important for baselining which is inherently connected to data cleaning). Although Akeem can also rely on high-resolution data from his battery asset, limitations in data storage and metering reliability will make data cleaning more challenging due to the reduced volumes of historical data available for gap filling for example. Thus, a DNO may view the hydro plant as a more 'market reliable' asset where flexibility and data provision are concerned (not necessarily in terms of its ability to deliver a successful flex service).This situation, involving data disparity and differences in control, can be deemed as an inequitable participation within a flex market as the varying asset specifications will afford Akeem and Jade different levels and methods of data cleaning before submitting data to the DNO. The level of asset and human resources needed for each party will also differ.

Having the DNO standardise data cleaning will level the participation field, especially where the handling of different resolutions, data volume, and asset types are concerned. Though not easy to get right or 'fair' in a competitive market, DNO management of data cleaning and processing will add standardisation but may raise further complex questions around data cleaning needs for different assets to validate flex services that we have yet to encounter.

5. Industry insight

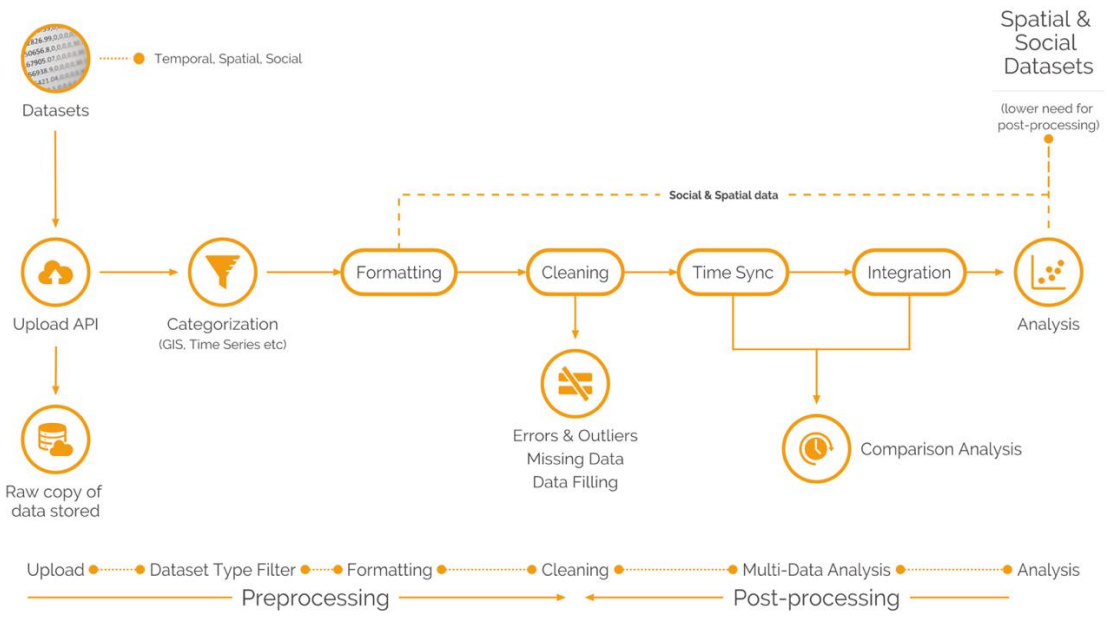
Data cleaning is not new to energy systems and balancing markets have their own procedures for the handling data from diverse assets at a national scale, potentially reducing resource needs from generators in terms of data cleaning. However, it is important to note that data cleaning is handled very differently and passively at these scales, where the lack of data is more of a hinderance to the data providers. As per the instruction and protocols for the Initial Settlement Run as outlined in Appendix 5.1 and 5.2 in the Overview of Trading Arrangements (2021) report by Elexon, missing data management is summarised as:

> *When incomplete data is submitted for an Initial Settlement Run, Section T1.4.5 of the BSC **[Balancing and Settlement Code]** states that the SAA **[Settlement Administration Agent]** should form an opinion on whether the data is 'substantially complete' before seeking instructions from BSCCo **[Balancing and Settlement Code Company]**. In practice, however, both BSCCo and the SAA prefer that BSCCo should take the lead in reaching decisions on these issues. For this reason, the SAA will inform BSCCo of missing or invalid data in all cases.*

Thus, the onus is on data providers to submit as complete as possible datasets to reach the appropriate settlement stages. However, it is unclear how these decisions on missing data are made from this report to determine what constitutes as 'substantially complete' data. DNOs in local flexibility markets may adopt a similar approach whereby procedures are in place for treating missing data, with simple data cleaning processes involved, but this may be at the risk of asset owners losing control of how their services are reflected data wise within settlement stages.

6. First Suite of LEO Tools

Data cleaning tools in LEO were first discussed within the first version (Internal access) of this report and much of the report focused on methodologies that were developed for the cleaning of timeseries data. The following paragraphs give a summary of this work. The figure below is shown as a refresher of how LEO's data are handled in terms of pre- and post-processing.



The Methods Flow Diagram above shows how datasets transition through the Pre- and Post-processing stages.

To effectively analyse data within LEO, data cleaning and quality checks are needed to ensure accurate learnings. Data tools have been developed to process data from a 'Formatting' stage of cleaning to one of 'Error Detection' where missing data and outliers are screened. Once errors have been found, various solutions are applied to clean the data from its raw format. In LEO, we have developed a 'multi-label classification' methodology to clean data as well as provide proper metadata on the cleaning techniques applied. This technique involves an algorithm which scans through each 'row' in a timeseries dataset, applying the multi-label classification method which mimics an 'on/off' status depending on the errors found. Effectively, each data point is tagged for the 'Errors' and 'Solutions' applied, thereby providing clear data provenance. See the first version of this report[1] for further detail or contact Project LEO for further information. However, this report was largely conceptual and further work post-report publication was done to create Python algorithms to implement this scheme as discussed in later reports, including this one.

6.1. Dash Tools for open access

Dash by Plotly is a unique suite of open-source libraries that has allowed us at LEO to build user-friendly and highly interactive data cleaning tools. Dash strips away the gritty code running the data cleaning, allowing a completely unfamiliar user the ability to clean their data from anywhere and through their web browser of choice. In LEO, we will build these tools (only progress to-date is reported on here) for both internal and external stakeholders to easily access. The packages and open-sourced libraries running in the backend are hosted using Heroku, enabling uses to access these tools through a URL. Previous tools such as the Time Syncing Tool were built on Jupyter Notebooks, but the use of this tool is not inherently obvious to the average user as much of the code running the analysis is 'exposed' and needs to be configured. Furthermore, the need and use of this tool with LEO did not warrant migration to a Dash web-based application.

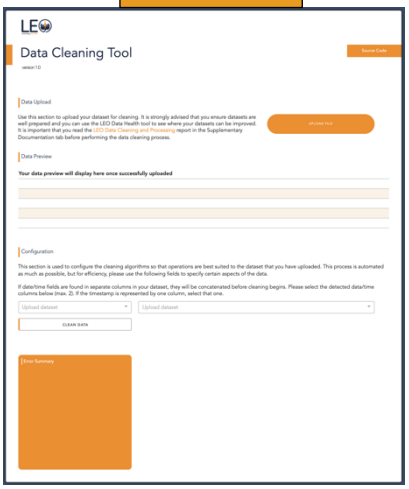
This report largely focuses on the two current data cleaning tools, the Data Cleaning Tool and the Data Health Scan, both of which are web-based applications that have been developed to support analysis in LEO with both publicly available version developed since the previous version of this report[1].

6.2. Data Cleaning Tool

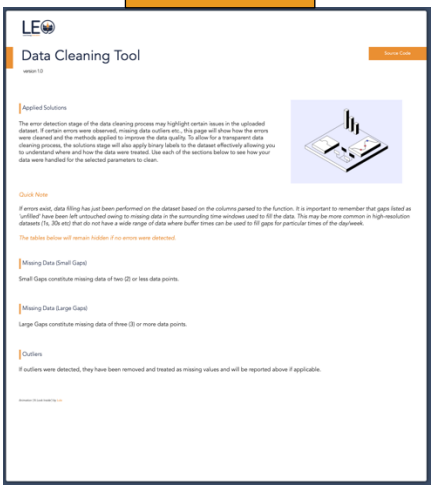
Preliminary work with the Data Cleaning Tool was discussed in the previous version of this report[1], but further modifications, including a launch to beta testers, have since taken place and are of focus within this report. Below, we have the three main tabs (6 in total): Overview, Errors Report, Solutions Report.

Overview

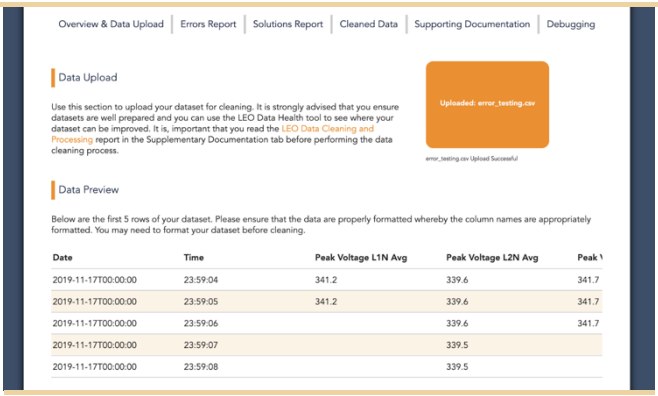
Errors Report

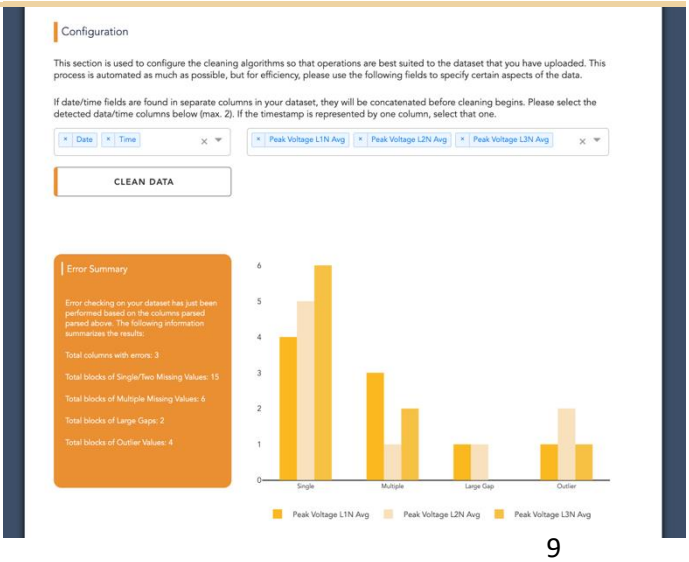


Solutions Report

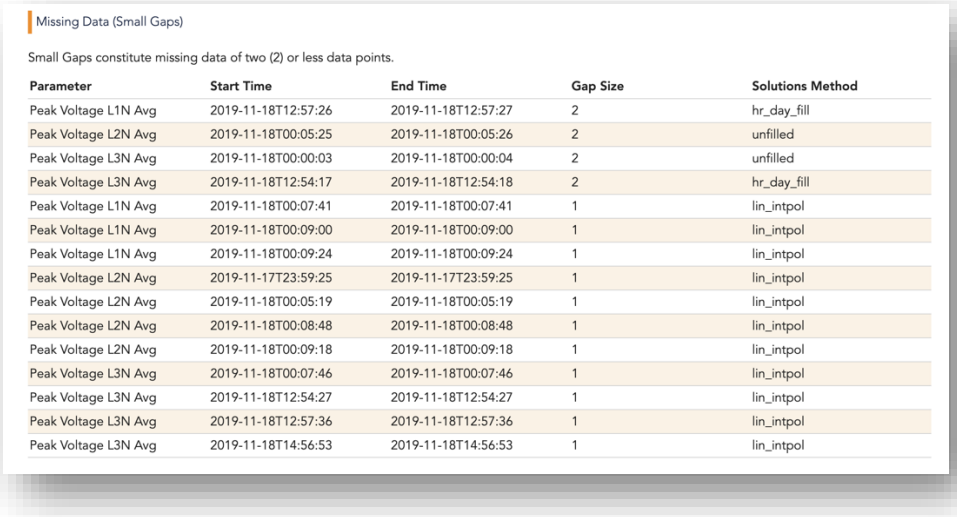


These tabs guide the user in the data cleaning process and sit alongside three other tabs : the Cleaned Data tab for data downloading (post-processing), the Supporting Documentation tab for further information for the cleaning behind the scenes, and the Debugging tab for reporting any issues. Below are example screen captures of how the pages look once a dataset has been uploaded by a user and a cleaning scan has been run. All tables and charts are very interactive and give users a unique opportunity to quickly access and visualise their data.





Tabs such as the *Solutions Report* and *Cleaned Data* (both seen below) allow the user to apply pre-set cleaning methods (largely various interpolation methods depending on the errors within the data) and download the dataset in its raw + cleaned format, or simply, only the cleaned data.

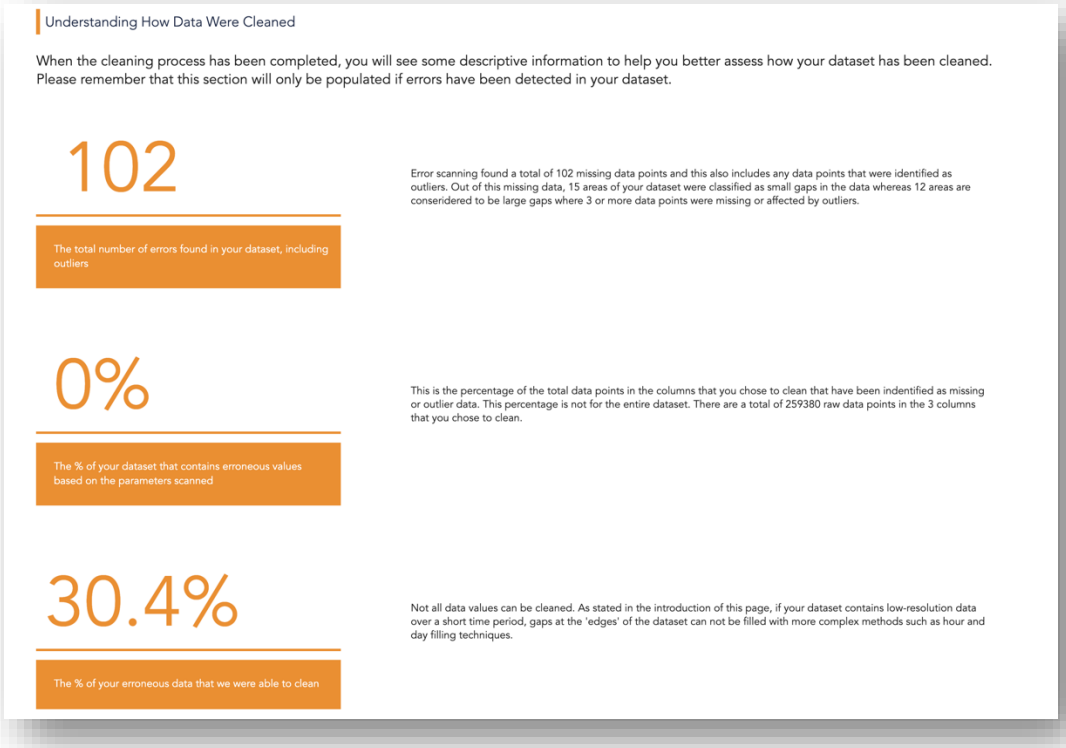


Missing Data (Small Gaps)

Small Gaps constitute missing data of two (2) or less data points.

Parameter	Start Time	End Time	Gap Size	Solutions Method
Peak Voltage L1N Avg	2019-11-18T12:57:26	2019-11-18T12:57:27	2	hr_day_fill
Peak Voltage L2N Avg	2019-11-18T00:05:25	2019-11-18T00:05:26	2	unfilled
Peak Voltage L3N Avg	2019-11-18T00:00:03	2019-11-18T00:00:04	2	unfilled
Peak Voltage L3N Avg	2019-11-18T12:54:17	2019-11-18T12:54:18	2	hr_day_fill
Peak Voltage L1N Avg	2019-11-18T00:07:41	2019-11-18T00:07:41	1	lin_intpol
Peak Voltage L1N Avg	2019-11-18T00:09:00	2019-11-18T00:09:00	1	lin_intpol
Peak Voltage L1N Avg	2019-11-18T00:09:24	2019-11-18T00:09:24	1	lin_intpol
Peak Voltage L2N Avg	2019-11-17T23:59:25	2019-11-17T23:59:25	1	lin_intpol
Peak Voltage L2N Avg	2019-11-18T00:05:19	2019-11-18T00:05:19	1	lin_intpol
Peak Voltage L2N Avg	2019-11-18T00:08:48	2019-11-18T00:08:48	1	lin_intpol
Peak Voltage L2N Avg	2019-11-18T00:09:18	2019-11-18T00:09:18	1	lin_intpol
Peak Voltage L3N Avg	2019-11-18T00:07:46	2019-11-18T00:07:46	1	lin_intpol
Peak Voltage L3N Avg	2019-11-18T12:54:27	2019-11-18T12:54:27	1	lin_intpol
Peak Voltage L3N Avg	2019-11-18T12:57:36	2019-11-18T12:57:36	1	lin_intpol
Peak Voltage L3N Avg	2019-11-18T14:56:53	2019-11-18T14:56:53	1	lin_intpol

Solutions Report Sample Data Table



Understanding How Data Were Cleaned

When the cleaning process has been completed, you will see some descriptive information to help you better assess how your dataset has been cleaned. Please remember that this section will only be populated if errors have been detected in your dataset.

102

The total number of errors found in your dataset, including outliers

Error scanning found a total of 102 missing data points and this also includes any data points that were identified as outliers. Out of this missing data, 15 areas of your dataset were classified as small gaps in the data whereas 12 areas are conseridered to be large gaps where 3 or more data points were missing or affected by outliers.

0%

The % of your dataset that contains erroneous values based on the parameters scanned

This is the percentage of the total data points in the columns that you chose to clean that have been indentified as missing or outlier data. This percentage is not for the entire dataset. There are a total of 259380 raw data points in the 3 columns that you chose to clean.

30.4%

The % of your erroneous data that we were able to clean

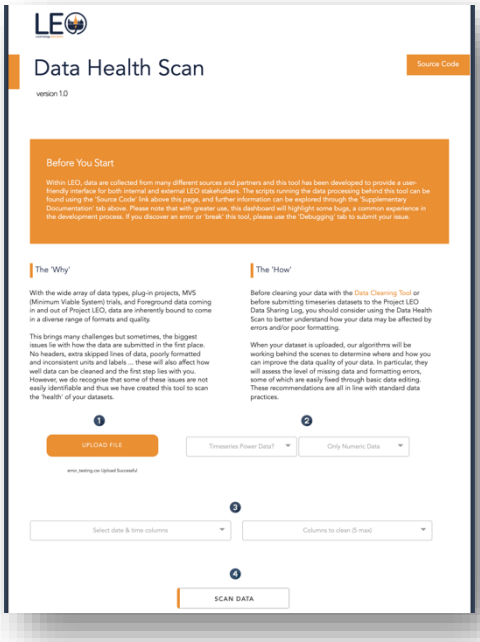
Not all data values can be cleaned. As stated in the introduction of this page, if your dataset contains low-resolution data over a short time period, gaps at the 'edges' of the dataset can not be filled with more complex methods such as hour and day filling techniques.

Cleaned Data Report Sample Cleaning

As seen above, the tool gives a comprehensive overview of the submitted dataset and how it was cleaned, including the limitations involved in the cleaning methods. Users are then able to download their cleaned data.

6.3. Data Health Scan

We have added another useful tool, the LEO Data Health Scan, that will fall within the full suite of cleaning dashboards. This tool will allow users to scan the 'health' of their datasets before performing any data cleaning steps. The Data Health Tool will ingest datasets provided by a user and then display interactive gauges (as seen on the following page) that will report key metrics such as the percentage of missing data. Note, automation can only cover so much when it comes to complex datasets and the onus is on the user to ensure that they bring datasets up to certain widely accepted standards before parsing them to our tools.



Sample reporting, all on one tab

Users can get an idea of the missing data, outliers, and possible time gaps in their timeseries data using this tool. This is helpful in knowing the level of pre-processing that may be needed before data cleaning.

6.4. Beta Testing

The Data Cleaning Tool has been exercised by a small subset of LEO data users and partners testing both functionality and utility since February 2022. Results and feedback will be fed directly into the development of the tool, however, with the recognition that this tool will inherently be limited in functionality due to resource constraints. For instance, there is one particular asset whose metering formats datasets transposed to the usual row *x* column formatting for timeseries data. The tool is unable to handle the ingestion of such formats and although configuration can be added to the tool to allow for processing of these types of data, there will be a limit to the ability to automatically handle the diversity in data, implying that data providers will need to have a level of pre-processing to meet regularly accepted data standards and formatting. Issues like this one can only be teased out through testing to spot holes in development and room for improvements.

7. Where Next?

With Project LEO in its final year, data cleaning will move out of development stage and more into an open space where results and tools are disseminated in future workshops for key stakeholders. Work will continue with these web-based tools in line with user needs and other tools (cleaning or otherwise) will also be considered. As many internal and external stakeholders who prefer the flexibility and customisability of the backend scripts, we will also be making these scripts available through repositories and the global Python library, giving anyone across the world the ability to install our packages as open-sourced tools such as the pending *Power Clean* package in seconds.

'Local' can take on a whole new meaning in this regard as we work hard in LEO to fulfil our commitment to FAIR and open data management systems. Data management in LEO needs to keep replicability at its core to ensure that learnings can be effectively translated by *fast-followers* within other local energy systems.

We will also work outside the functional scope of these tools to address data cleaning in flex services. Who is responsible? How can standards ensure fair participation? What responsibility lies with the asset owners and what level of 'clean' data is needed for validation and baselining? Industry and internal conservations will help us to better understand these issues to feed into future reports and workshops.

All icons have been openly accessed from Flaticon