LEO Local Energy Oxfordshire

March 2021 | Version 2

Data Cleaning and Processing

Masaō Ashtine

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

Version:	2.0	Date:	31/03/2021
Workpack*:	WP4	Deliverable:	4.15
Reviewed by:	Rajat Gupta		
Date:	09/04/2021		
Signed off by:	David Wallom		
Date:	09/04/2021		

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[1]

This report will highlight work that has been done on Project LEO's tools for data cleaning and quality control since the first version of this report was published in March 2020. This update will briefly touch on 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. The following sections will also focus on 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, 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. First Implementation

Data cleaning tools in Project LEO (LEO henceforth) were first discussed within version 1.0 of this report and much of the report focused on methodologies that were developed for the cleaning of timeseries data. The following paragraphs will give a brief summary of this work, but further details are out of the scope of this update report.

[1] *Note, 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. All proposed data platforms will be separate from the Integrated Land Use Mapping tool that has also been developed in LEO.*

With the wide array of data types, plug-in projects, MVS trials, and Foreground data coming in and out of Project LEO, data are inherently bound to come in a diverse range of formats, quality and access. The experimental nature of an MVS trial can produce data from many different equipment types (temporary monitoring equipment etc.), each having their own software and formats with the potential for missing data. Data, particularly timeseries data in any format (such as CSV), need to be investigated and cleaned thoroughly to maximise information that can be drawn from these datasets, without any erosion of data quality or the injection of unintended bias.

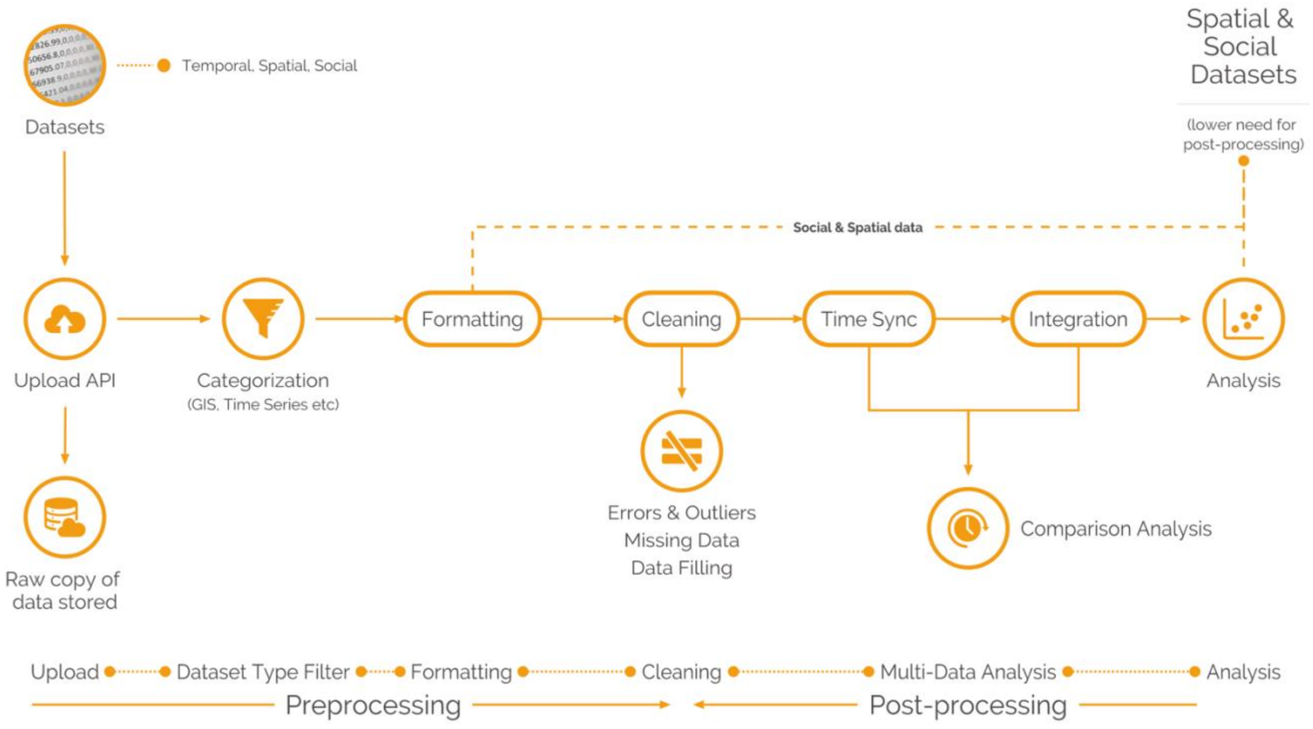
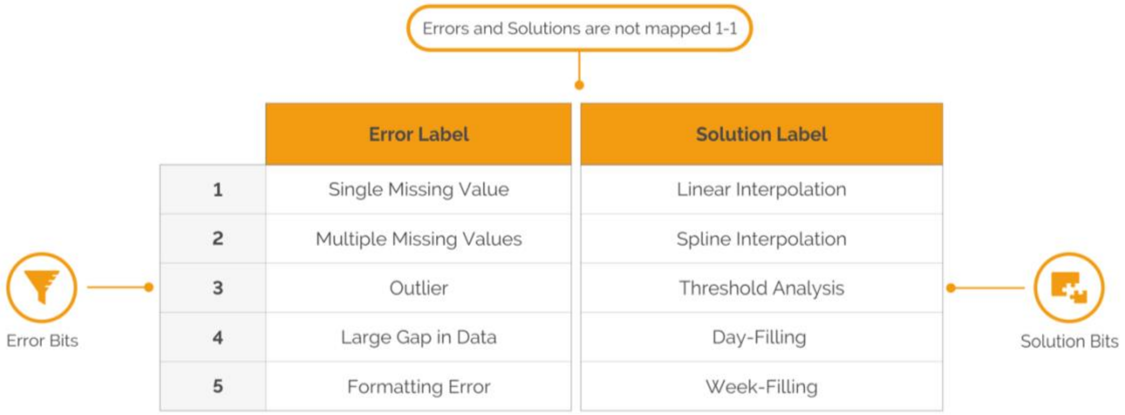


Figure 1. The Methods Flow Diagram above shows how datasets transition through the Pre- and Post-processing stages.

In order 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 listed in the table below. Effectively, each data point is tagged for the 'Errors' and 'Solutions' applied, thereby providing clear data provenance.



Errors and Solutions are not mapped 1-1
Error Bits
Solution Bits

	Error Label	Solution Label
1	Single Missing Value	Linear Interpolation
2	Multiple Missing Values	Spline Interpolation
3	Outlier	Threshold Analysis
4	Large Gap in Data	Day-Filling
5	Formatting Error	Week-Filling

Table 1. Error and Solution Bit labelling system that will be applied within LEO's data cleaning

The previous report describes this methodology in depth, providing use cases to demonstrate how data cleaning can be performed. However, this was largely conceptual and further work post-report publication was done to create Python algorithms to implement this scheme.

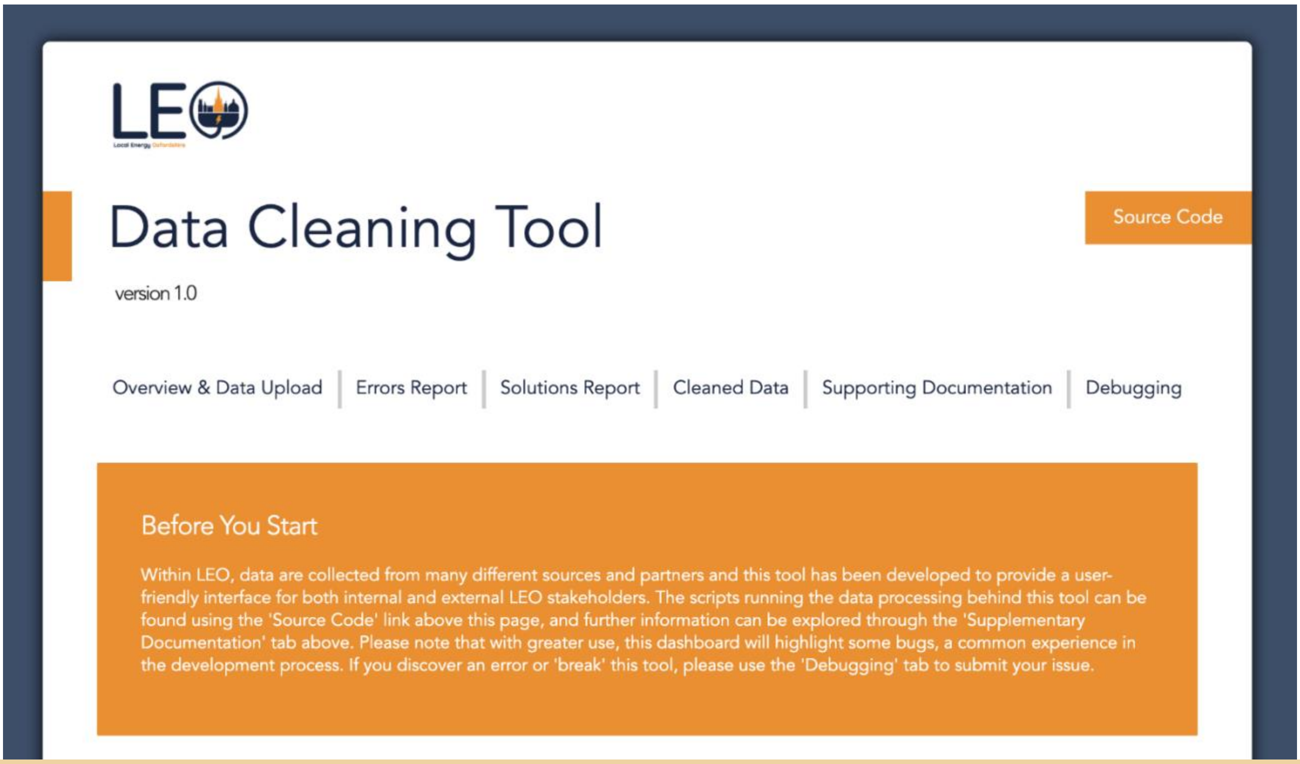
4. A New Suite of Tools

The multi-label classification scheme for error detection and data cleaning was translated into a Python module with various scripts and algorithms that allowed a skilled user the ability to comprehensively clean small and complex datasets alike. These well-documented scripts can be found on the LEO Data Repository, but they presented a new challenge … accessibility. With one script (of many) having circa 1,500 lines of Python code, these scripts, though powerful tools, did not allow the average participant in LEO access to our data cleaning tools. This is counterproductive to LEO's goals in data management as the project models itself as an accessible springboard for other local energy systems and their agents. The following sections will then detail the main improvements to transitioning existing cleaning tools to online and user-friendly dashboards, transforming how data can be interacted with through open-source platforms.

4.1. Enter Dash

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 will be hosted on a University of Oxford Virtual Machine, enabling uses to access these tools through a URL such as '*project-leo-data.co.uk'* for instance (TBC).

4.2. Main Data Cleaning Dashboard



Data Cleaning Tool

version 1.0

Before You Start

Within LEO, data are collected from many different sources and partners and this tool has been developed to provide a user-friendly interface for both internal and external LEO stakeholders. The scripts running the data processing behind this tool can be found using the 'Source Code' link above this page, and further information can be explored through the 'Supplementary Documentation' tab above. Please note that with greater use, this dashboard will highlight some bugs, a common experience in the development process. If you discover an error or 'break' this tool, please use the 'Debugging' tab to submit your issue.

Figure 2

Above, we have the top section (*Overview* tab) of the LEO Data Cleaning Tool, built on Dash (not yet deployed for wider use at time of this report). As show, this clean interface provides a responsive portal for accessing the tools and information needed for cleaning datasets.

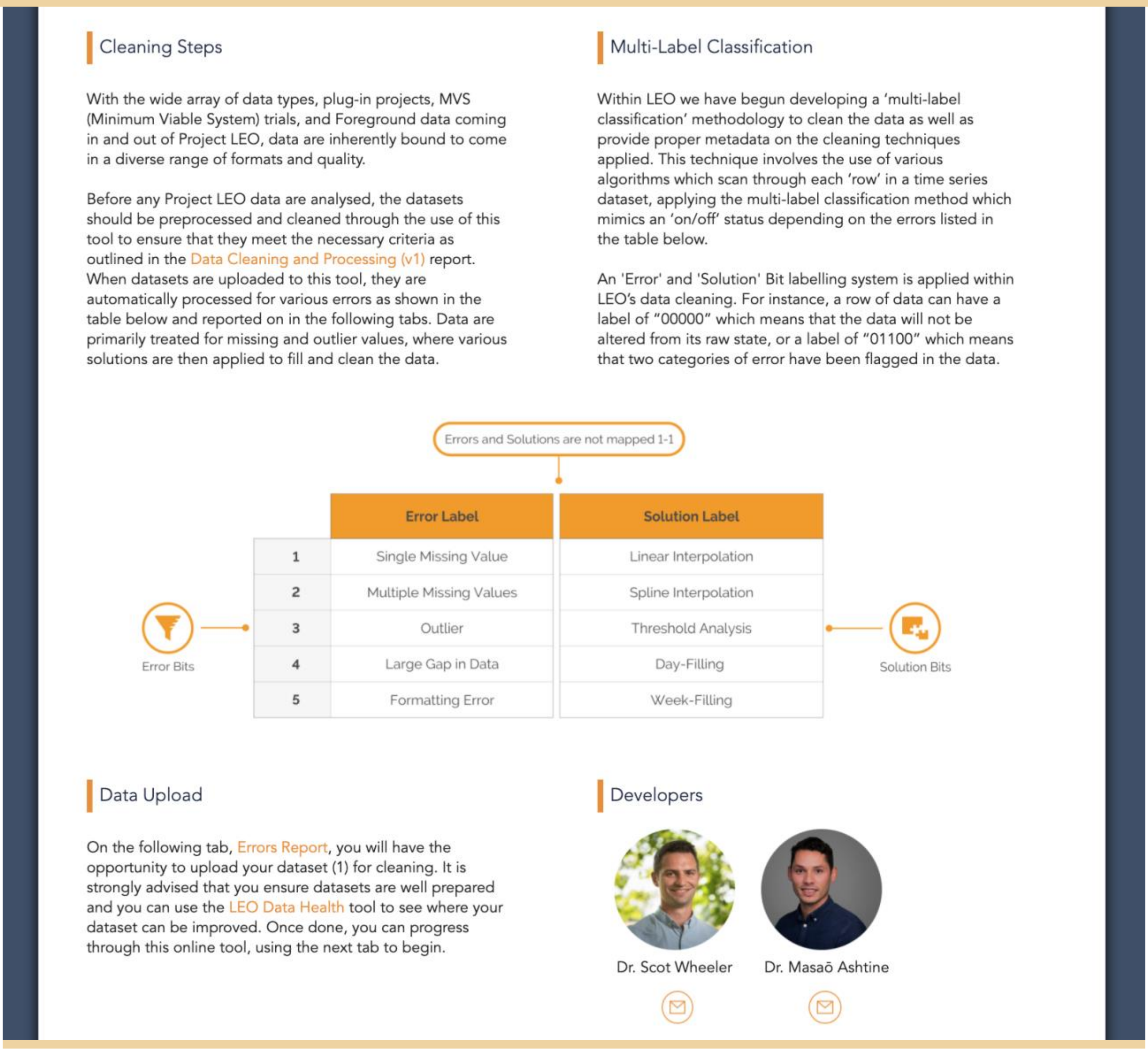


Figure 3

Cleaning Steps

With the wide array of data types, plug-in projects, MVS (Minimum Viable System) trials, and Foreground data coming in and out of Project LEO, data are inherently bound to come in a diverse range of formats and quality.

Before any Project LEO data are analysed, the datasets should be preprocessed and cleaned through the use of this tool to ensure that they meet the necessary criteria as outlined in the Data Cleaning and Processing (v1) report. When datasets are uploaded to this tool, they are automatically processed for various errors as shown in the table below and reported on in the following tabs. Data are primarily treated for missing and outlier values, where various solutions are then applied to fill and clean the data.

Multi-Label Classification

Within LEO we have begun developing a 'multi-label classification' methodology to clean the data as well as provide proper metadata on the cleaning techniques applied. This technique involves the use of various algorithms which scan through each 'row' in a time series dataset, applying the multi-label classification method which mimics an 'on/off' status depending on the errors listed in the table below.

An 'Error' and 'Solution' Bit labelling system is applied within LEO's data cleaning. For instance, a row of data can have a label of "00000" which means that the data will not be altered from its raw state, or a label of "01100" which means that two categories of error have been flagged in the data.

Errors and Solutions are not mapped 1-1

	Error Label	Solution Label
1	Single Missing Value	Linear Interpolation
2	Multiple Missing Values	Spline Interpolation
3	Outlier	Threshold Analysis
4	Large Gap in Data	Day-Filling
5	Formatting Error	Week-Filling

Data Upload

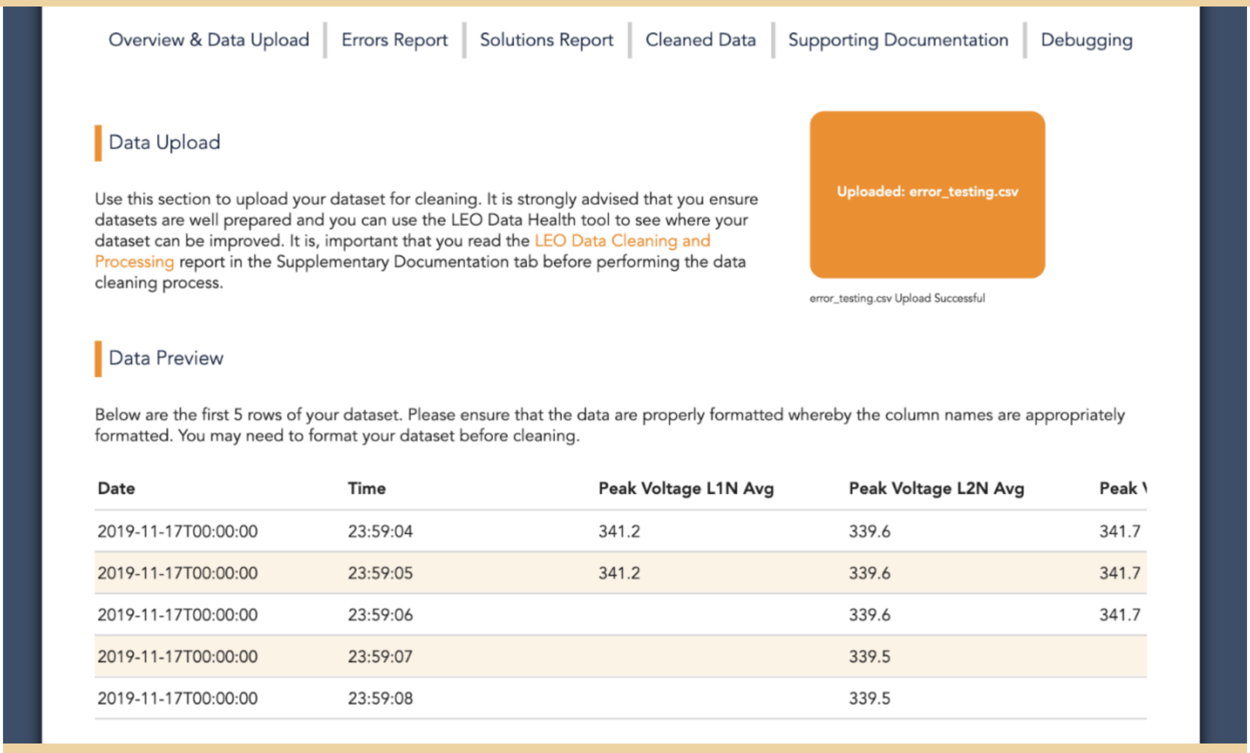
On the following tab, Errors Report, you will have the opportunity to upload your dataset (1) for cleaning. It is strongly advised that you ensure datasets are well prepared and you can use the LEO Data Health tool to see where your dataset can be improved. Once done, you can progress through this online tool, using the next tab to begin.

Developers

Dr. Scot Wheeler

Dr. Masaō Ashtine

Above, we can see another section of this landing page that briefly describes how datasets are cleaned and where further information can be accessed. In the *Data Upload* section, we present another important tool, the LEO Data Health Tool, that will fall within the full suite of cleaning dashboards. This tool will allow users the ability to scan the 'health' of their datasets before performing any data cleaning steps. 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. The Data Health Tool will ingest datasets provided by a user and then display interactive charts that will report key metrics such as the percentage of missing data or something as commonly overlooked as poorly structured columns and headers.



Data Upload

Use this section to upload your dataset for cleaning. It is strongly advised that you ensure datasets are well prepared and you can use the LEO Data Health tool to see where your dataset can be improved. It is, important that you read the LEO Data Cleaning and Processing report in the Supplementary Documentation tab before performing the data cleaning process.

error_testing.csv Upload Successful

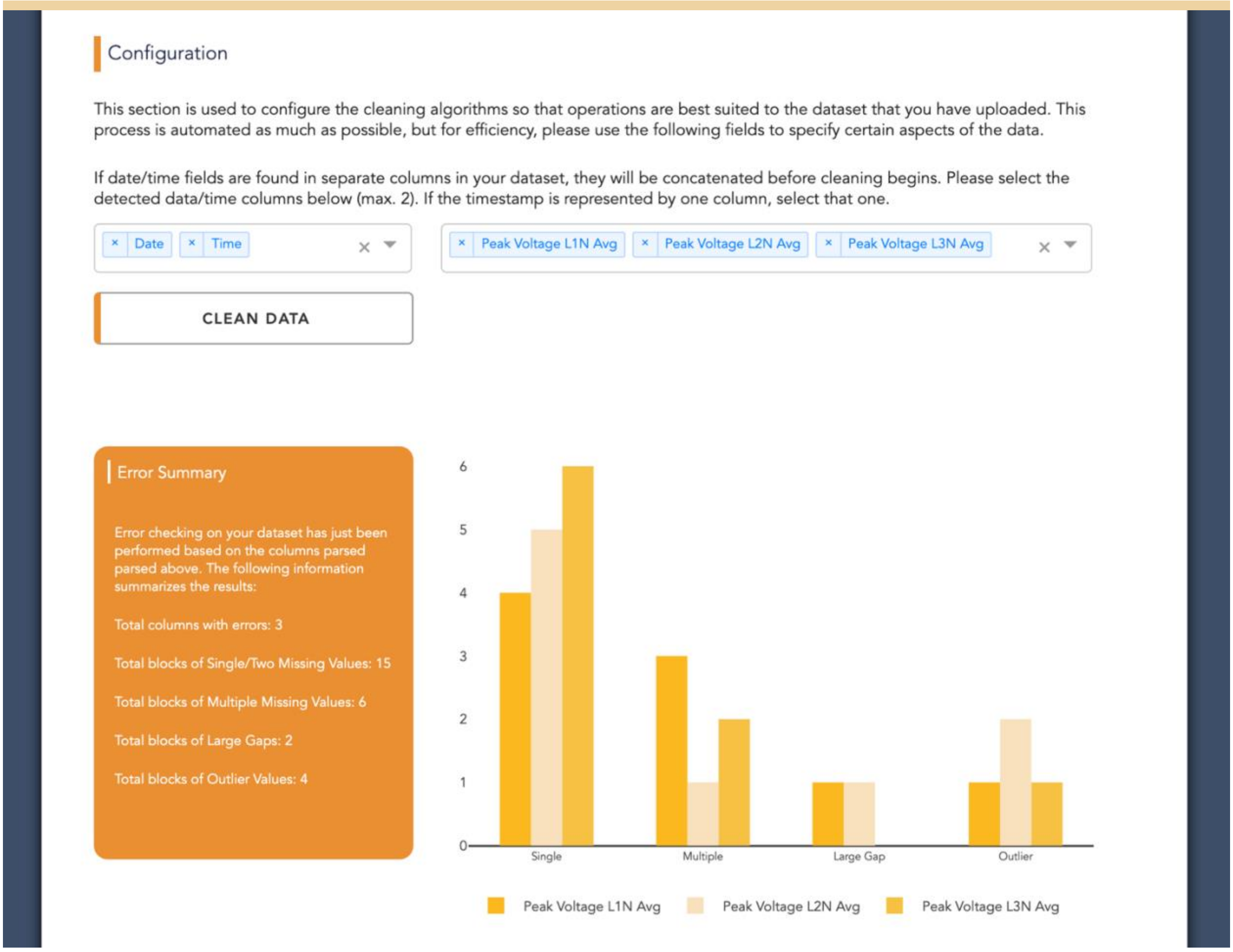
Data Preview

Below are the first 5 rows of your dataset. Please ensure that the data are properly formatted whereby the column names are appropriately formatted. You may need to format your dataset before cleaning.

Date	Time	Peak Voltage L1N Avg	Peak Voltage L2N Avg	Peak \
2019-11-17T00:00:00	23:59:04	341.2	339.6	341.7
2019-11-17T00:00:00	23:59:05	341.2	339.6	341.7
2019-11-17T00:00:00	23:59:06		339.6	341.7
2019-11-17T00:00:00	23:59:07		339.5	
2019-11-17T00:00:00	23:59:08		339.5	

Figure 4

Visually seeing some errors can only provide so much insight. After the dataset has been previewed, the user is presented with certain configuration fields which tell the cleaning algorithms where to focus on as seen below.



Configuration

This section is used to configure the cleaning algorithms so that operations are best suited to the dataset that you have uploaded. This process is automated as much as possible, but for efficiency, please use the following fields to specify certain aspects of the data.

If date/time fields are found in separate columns in your dataset, they will be concatenated before cleaning begins. Please select the detected data/time columns below (max. 2). If the timestamp is represented by one column, select that one.

Error Summary

Error checking on your dataset has just been performed based on the columns parsed parsed above. The following information summarizes the results:

Total columns with errors: 3

Total blocks of Single/Two Missing Values: 15

Total blocks of Multiple Missing Values: 6

Total blocks of Large Gaps: 2

Total blocks of Outlier Values: 4

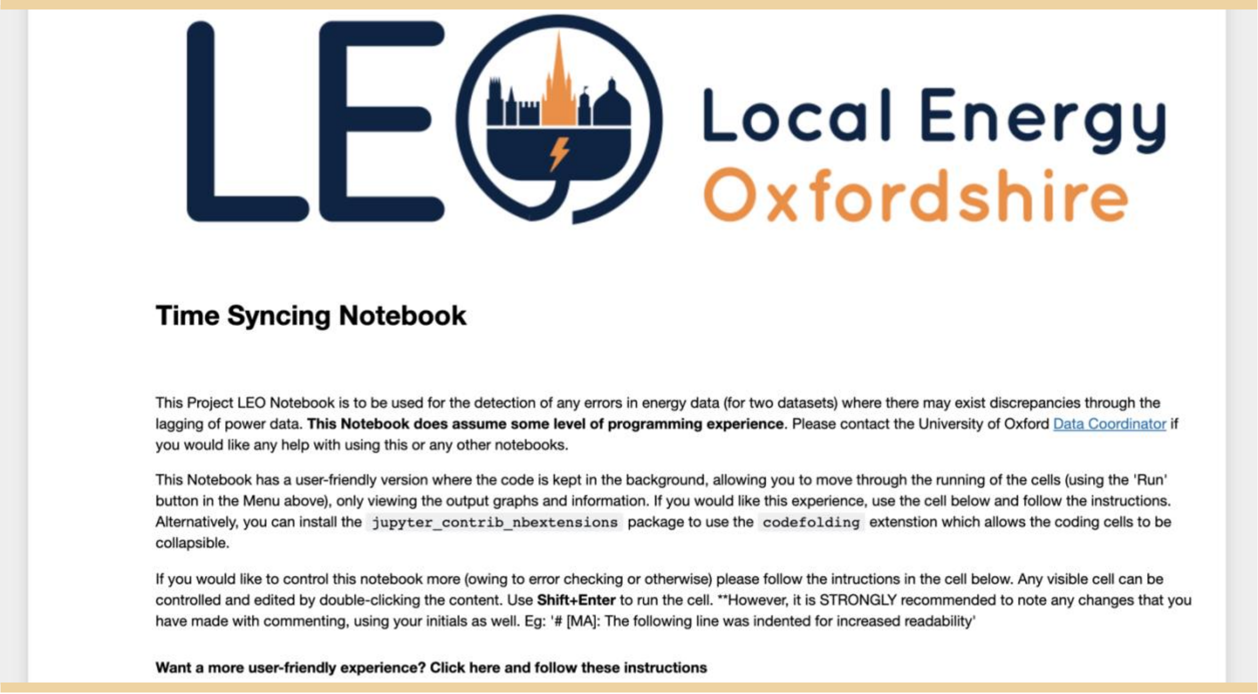
Figure 5

The user can select the columns that they would like to clean (max. 5 in version 1.0) and once they have entered in the necessary fields, clicking the *Clean Data* button will initialise the algorithms on the server to process the dataset. This tab is solely focused on the detection of errors within the dataset which are then displayed in the *Error Summary* section and subsequent interactive graph. As you can see, the user is less involved with the 'behind-the-scenes' and much more involved with interacting with their dataset through this cleaning interface. For each column cleaned, the graph will be automatically updated to show the magnitude of various error categories in the dataset (in this example, we can see that the *error_testing.csv* dataset had more *Outlier* gaps in the *Peak Voltage L1N Avg column*, and no large gaps (more than 10 consecutive missing data points) in the *[…] L3N Avg* data column.

Later tabs such as the *Solutions Report* and *Cleaned Data* tabs will allow the user the ability 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. The *Cleaned Data* tab will also provide users an interactive space where they can see the clean versus raw data to visualise how the data were treated.

4.3. Post Processing Tools

To complete the suite of data cleaning and data quality tools in LEO, other dashboards such as the Time Sync Tool (as seen below) will be deployed in a similar fashion to the Data Cleaning Tool.



Time Syncing Notebook

This Project LEO Notebook is to be used for the detection of any errors in energy data (for two datasets) where there may exist discrepancies through the lagging of power data. **This Notebook does assume some level of programming experience**. Please contact the University of Oxford Data Coordinator if you would like any help with using this or any other notebooks.

This Notebook has a user-friendly version where the code is kept in the background, allowing you to move through the running of the cells (using the 'Run' button in the Menu above), only viewing the output graphs and information. If you would like this experience, use the cell below and follow the instructions. Alternatively, you can install the `jupyter_contrib_nbextensions` package to use the `codefolding` extenstion which allows the coding cells to be collapsible.

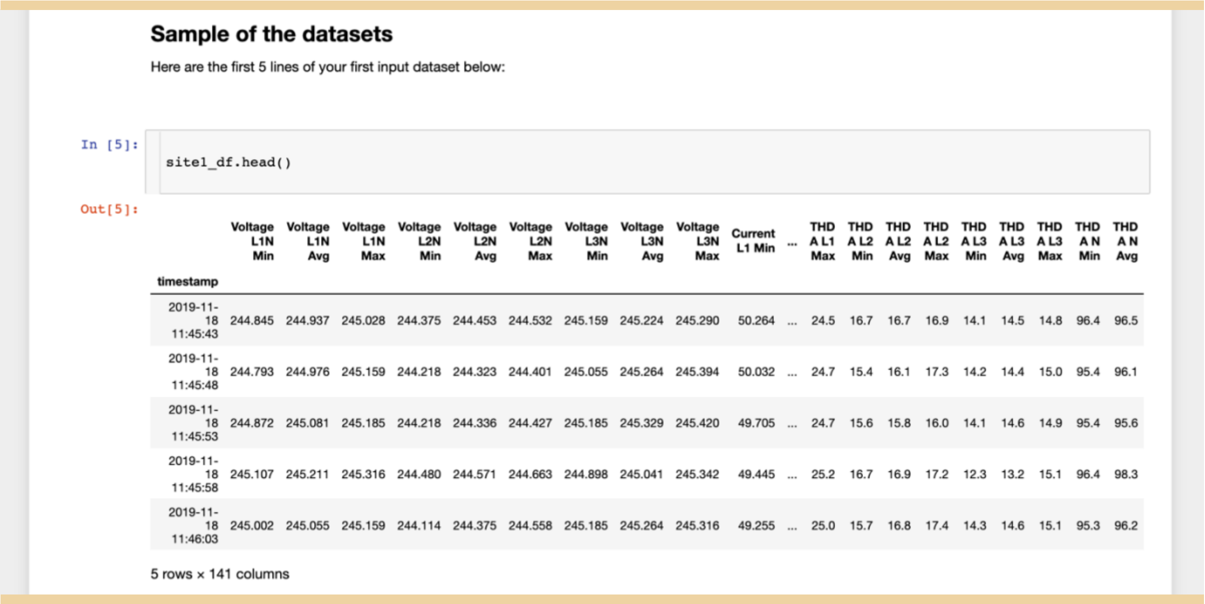
If you would like to control this notebook more (owing to error checking or otherwise) please follow the intructions in the cell below. Any visible cell can be controlled and edited by double-clicking the content. Use **Shift+Enter** to run the cell. **However, it is STRONGLY recommended to note any changes that you have made with commenting, using your initials as well. Eg: '# [MA]: The following line was indented for increased readability'

Want a more user-friendly experience? Click here and follow these instructions

Figure 6

This bespoke tool allows users, after data have been cleaned, to compare data from an asset (e.g.: a battery at a prosumer site) and from a substation to automatically determine if there are any lags in power measurements owing to errors in how the data are recorded and metered.

The last line in the image alludes to the fact that this tool is not fully accessible to all. Being built on a Jupyter Notebook, 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. Though developed as a middle ground interface for improved access, this tool will be transformed into an online dashboard where all a user needs are a URL, data, and web-browser.



Sample of the datasets

Here are the first 5 lines of your first input dataset below:

```
In [5]:  site1_df.head()
```

Out[5]:

timestamp	Voltage L1N Min	Voltage L1N Avg	Voltage L1N Max	Voltage L2N Min	Voltage L2N Avg	Voltage L2N Max	Voltage L3N Min	Voltage L3N Avg	Voltage L3N Max	Current L1 Min	...	THD A L1 Max	THD A L2 Min	THD A L2 Avg	THD A L2 Max	THD A L3 Min	THD A L3 Avg	THD A L3 Max	THD A N Min	THD A N Avg
2019-11-18 11:45:43	244.845	244.937	245.028	244.375	244.453	244.532	245.159	245.224	245.290	50.264	...	24.5	16.7	16.7	16.9	14.1	14.5	14.8	96.4	96.5
2019-11-18 11:45:48	244.793	244.976	245.159	244.218	244.323	244.401	245.055	245.264	245.394	50.032	...	24.7	15.4	16.1	17.3	14.2	14.4	15.0	95.4	96.1
2019-11-18 11:45:53	244.872	245.081	245.185	244.218	244.336	244.427	245.185	245.329	245.420	49.705	...	24.7	15.6	15.8	16.0	14.1	14.6	14.9	95.4	95.6
2019-11-18 11:45:58	245.107	245.211	245.316	244.480	244.571	244.663	244.898	245.041	245.342	49.445	...	25.2	16.7	16.9	17.2	12.3	13.2	15.1	96.4	98.3
2019-11-18 11:46:03	245.002	245.055	245.159	244.114	244.375	244.558	245.185	245.264	245.316	49.255	...	25.0	15.7	16.8	17.4	14.3	14.6	15.1	95.3	96.2

5 rows × 141 columns

Figure 7

Though useful to a programmer, interfaces such as this present certain limits for the average LEO user (*values shown here are from a raw dataset ingestion*)

5. Where Next?

Significant changes have been made to how partners and affiliated stakeholders in LEO can perform data cleaning and quality control. However, these tools are still underdevelopment and have not been fully rolled out. The Data Cleaning Tool will be completed whereby a deployed testing ground will allow as many people as possible the opportunity to 'break' and 'crash' the online dashboard, effectively giving the developers insight on improvements to be made. The *Debugging* tab for instance will enable users to upload screenshots of bugs and errors in the tool for developers to flag issues.

Work will continue on the Data Health Tool as well as the onboarding of post-processing analytical tools such as the Time Sync Tool to more user-friendly dashboards. However, there are many internal and external stakeholders who prefer the flexibility and customisability of the backend scripts themselves! Thus, we will also be making these scripts available through 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.

POWER CLEAN
AN OPEN-SOURCE TOOL BY PROJECT LEO

'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. Thus, the future design of our analytical tools will involve a careful balance of open-access tools, trending and widely adopted software for data analysis, and user-friendly interfaces that increase a project's engagement with its data.