The Need for Tidiness in Energy Data Analysis

Joseph Scott

*School of Engineering
Newcastle University*

Newcastle upon Tyne, United Kingdom
j.scott9@ncl.ac.uk

Dr Neal S. Wade

*School of Engineering
Newcastle University*

Newcastle upon Tyne, United Kingdom
neal.wade@ncl.ac.uk

Dr Johnson Fernandes

Equiwatt Limited

Gateshead, United Kingdom
johnson@equiwatt.com

Abstract—The move to Smart Grids, has turned up a wealth of data improving the ability to understand various aspects of the network. Usage of this data offline (post trial) can be hindered by the structure of the data. Tidy principles result in a systematic structure beneficial for analysis. We argue tidiness of data is being overlooked in other work to organise energy data. The repositories of several power systems projects are examined to understand how the data is structured. The findings are summarized and recommendations made for improvements in tidiness. Future works in the area are suggested.

Index Terms—energy-data, open-data, data-engineering, smart-grid, tidy-data

I. INTRODUCTION

Several large trials have or are publishing data, this paper argues on the need for systematic organization or tidiness [1]. The process of data Extract-Transform-Load (ETL) [2] into a form suitable for analysis is the beginning of any study. This may be a one off task or incremental but researchers will be doing some stages and degree of ETL. Understanding which data to extract from the dataset, how to transform and manipulate it for loading into the downstream software can be hindered by the organization. Tidiness is concerned with organization to reduce redundancy, make illegal relations impossible and permit efficient analysis. Understanding the organization comes from project materials and helper ‘readme’ files but the repetition of ETL for each dataset frustrates the analysis process.

This paper suggests application of ‘tidy data’ principles which give a consistent structure to data. Several similar concepts exist in energy systems, the common information model (CIM) improves interoperability by defining a family (ontology) of suitable concepts for various applications. Metadata¹, such as the CIM also help understand a dataset’s contents at a glance. But tidy data concerns a schema: the structure in how rows and columns fit together across several groupings of data. This paper and its topic of tidy data is not concerned with broader requirements in ETL such as outlier removal nor any particular file format or database and their associated performance. The merits of any method or tool are highly domain and scenario specific.

Joseph Scott is supported by the European Regional Development Fund for North East England and Equiwatt Limited.

¹Loosely in this context ‘data about the data’.

A. Organization

The remainder of the paper organized as follows: In II the concept of tidiness is introduced from existing literature and compared with current movements and literature for energy data. Secondly, III tours energy systems projects including smart-metered, electric vehicle and disaggregation data. Their schemas, where provided, are discussed and compared with the provided dataset in order to evaluate their tidiness. In IV recommendations based on the patterns found are considered, based on the solutions seen in other datasets and the concepts of tidiness. Finally in V conclusions are drawn and suggestions for future work is identified.

II. TIDY DATA

The first subsection introduces the tidiness outlining its origin, why it is beneficial and signs of messy data. The second subsection, outlines a recent work by a UK taskforce to improve energy data through specific and actionable means. We argue the taskforce’s scope is not considering this area and that although research into other aspects of ETL energy data exist, concepts relating to tidiness are absent in the application domain of energy researchers.

A. Basis in Tidy Data

Tidy data is described by Wickham [1] as:

A re-framing of Codd’s 3rd normal form but with constraints reframed in statistical languages ... (the) principles of tidy data provide a standard way to organize data values within a dataset...

- ‘Each variable forms a column’ (i.e. any measurement electrical or otherwise.)
- ‘Each observation forms a row’ (i.e. associated measurements collected e.g. timestamp, voltage and current.)
- ‘Each observational unit forms a table’ (i.e. a set of time series of measurements.)

Wickham’s tidiness definition is justified because a consumer of the data can more readily extract subsets by filtering for observations and selecting desired variables from an observational set. The assumptions also carry into efficient development of programs consuming data, as Wickham says in section 4, arguing that ‘messy tools’ consume or produce data not properly normalized. The disadvantage then with messy tools is the overhead to inter-operate both from the perspective

of the user and developer. In the *R* programming language, tidiness has spawned the *Tidyverse*² set of packages all for manipulation of tidy data.

Wickham states the common offending patterns against these rules:

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

These rules are applied to the examples in III to assess datasets tidiness. Clearly in an analysis, data will be denormalised for consumption but a tidy dataset has the hard work of separation.

B. Energy ETL and a Need for Tidiness

The UK Energy Data Task force [3], among several recommendations, outlines the need for data to be visible. In the report's second recommendation, 'Maximizing Value of Data', it is suggested that opening data is a step in achieving this recommended goal. Data should be 'Discoverable and Searchable and Understandable' with common 'Structures, Interfaces and Standards'. They argue this means accessing both a dataset's availability for usage but also making them 'easier to access and understand'. They go on to state:

The evolution of the energy sector, and the digital systems which support it, have resulted in a proliferation of different approaches to data and a range of incompatible data structures and interfaces. In order to realise the benefits of a smart, flexible Energy System there is a need for the pragmatic standardisation of data structures and interfaces to enable interoperability, improve collaboration and allow innovation to be scaled up effectively.

However this is a level above the concern of tidiness, the recommendation above deals with the ecosystem and metadata. It is argued that tidiness then is a part of this recommendation, improving access in practice to the datasets. At the time of writing a further collection and curation of data best practice³ is available, containing 'actionable and specific' recommendation for datasets in terms of accessibility and barriers to that. This includes documenting a dataset's metadata detail by anticipating and exemplifying issues which arise from privacy or sensitivity. Schemas (and therefore tidiness) are mentioned as existing, with some minor notes on dealing with particular data types e.g. ISO 8601 usage for date-times.

If publishing a schema is an assertion about a closed world, usage of a standardized domain model eases burden and development of that standard makes this flexible over inevitable change.

ETL and data processing strategies have received attention in the energy domain, in [4] methods are surveyed for missing

value imputation, evaluating the error of various methods. Others review the trial-family methodological differences, such as [5] which reviews smart-meters and [6] disaggregation datasets. Each describe the domain specific choices for that family of trials, whereas tidy data outlined by Wickham and as discussed in this paper refers exclusively to the organization of data semantically, so that it may consistently be accessed. We argue this thinking ought to be applied to all these datasets, with a domain model to permit access to much more data. If the recommendations quoted by the UKEDTF are to be followed, this must be acknowledged.

III. TOURING TIDINESS IN ENERGY DATA

This section tours the dataset outputs from trials and examines their tidiness. The naming is deliberate, the datasets were already accessed for analysis purpose so it is misleading to call this a review. The first area of non-intrusive load monitoring (NILM) uses disaggregation to identify devices present in a aggregate metered signal. These trials involve monitoring circuit sub-branches, meters and larger appliances themselves and then allocating power to the set of appliances. Several datasets are available, all alike in framing but varying in methodological aspects: measurement equipment, sampling period, number of buildings and amount/type of appliances monitored. Supervised and unsupervised analysis are used with neural networks, hidden markov models, combinatorial optimization and change-point clustering all common approaches to disaggregation. For an overview of datasets see [7], although this section will focus the work around the NILM-ToolKit (NILMtk)⁴ which provides a schema and 'glue code' in order to compare the several datasets available. See Hart [8] for an introduction to features considered within this area.

Secondly, a group of larger UK trials, including Smart Metering in London (SMiL)⁵, My Electric Avenue (MEA)⁶ and (the) Customer Led Network Revolution (CLNR)⁷ are considered. Of the three only SMiL's dataset resides in a dedicated archive, (the) City of London's, while CLNR and MEA are hosted on their respective websites (see footnoted links). These datasets are introduced in order of complexity: SMiL, MEA and CLNR, the last of which is the only with a diagrammatic schema document.

A. Disaggregation for Non-intrusive Load Monitoring

Disaggregation has been an active research area since Hart's beginnings. As a result several datasets are available each with diversity in scale, duration, features metered and organization. In response, the NILM toolkit [9] implements a library of functions and uses a metadata-schema to parse the differing files into common form for analysis. This metadata-schema used is published by Kelly et al [10] which presents a Uniform Modeling Language (UML) model of the NILM trials. Tidiness principles can be seen in *Figure 1* of [10] of the paper

²<https://www.tidyverse.org/>

³<https://es.catapult.org.uk/guides/energy-data-best-practice-guidance/>

⁴<https://github.com/nilmtn/nilmtn>

⁵<https://data.london.gov.uk>

⁶<http://myelectricavenue.info/>

⁷<http://www.networkrevolution.co.uk/>

or reproduced online⁸. The library implements each dataset adapting it for consumption by a common library designed to work with data implemented according to the schema. The UML model does not directly describe the data, this is what makes it a meta-schema. The datasets are presented in a tidy form with variables forming observations found together and a consistent tidy pattern to access the data.

The NILM community has centered on separating metadata from data, recognizing the aspects of the trial which differ and provide context to the data. Additionally each trial has differing equipment and appliance mix due diversity in location and implementation decisions. Where the variables match (i.e. units/dimensions are compatible) there is a comparison between the sets, typically the current or power. Organization of this is a hurdle and the NILM community benefits greatly from unification of the datasets.

B. Large Scale Trials

These large trials gathered bodies of data centered around their research questions. They all feature sets of participants but may not have involved all participants in all aspects of the trials. In terms of tidy principles, various sections of trials may gather unique observations or require several further tables to express differing observations. The set of trials have observations of fundamentally differing entities or observations.

1) *Smart Metering in London*: This dataset provides smart metering data between 2011 and 2014, researching the link between behavior/attitudes and energy usage. The data collects a representative sample of the Greater London Area totaling 5500 customers. A thousand of these were placed on time of use tariffs, while others acted as a control. Socio-economic behavioral classifications are included forming part of the behavioral aspects. The project notes they must assume units of kWh for the samples. The official source provides the following files:

- ‘low-carbon-london-data.zip’ — The dataset files containing CSVs
- ‘Tariffs.xlsx’ — Contains the time of use tariffs.

The columns present in the data zip indicate it is not tidy. In each row multiple observations are made including: ‘KWH/hh (per half hour)’, ‘Acorn’ and ‘Acorn_grouped’. For this data to be tidy it should be separated into multiple tables for it’s observations. It is argued distinct observations are participants, acorn groups and smart-metered samples. The remaining column is the trialist group, defining membership of time of use tariffs. According to the repository’s description, time of use applies to the subset only over the 2013 period. It is argued that this variable should go with participant observations.

Kaggle is a collaborative data analysis site hosting repositories of data and kernels. Organizers of Kaggle contests, to encourage contributions, will themselves make the problem and data accessible. This may include framing the problem, providing subject specific background, gathering additional

data (i.e. weather) and tidying the data. The following Kaggle project⁹ tidies and augments this dataset with weather data, it also includes questionnaire results.

2) *My Electric Avenue*: My Electric Avenue investigated electric vehicle (EV) behaviors in two trials. The first, the ‘social’ trial gathers data on users trips and charging times. In the second, clusters of vehicles were metered as before but in addition charge curtailment was applied at based on the transformer load. Users willingness to provide flexibility in EV charging was monitored by a user facing app. Based on the trial, conclusions were drawn as to the extent EV uptake would affect substation headroom. The following files are provided:

- ‘Clusters.csv’ — The file defining the clusters including social trialists clusters.
- ‘EVChargeData.csv’ — A derived series of charge events from the EVs.
- ‘EVTripData.csv’ — A derived series of trip events taken in EVs.
- ‘ICBData.csv’ — Metered time series file of vehicle telemetry.
- ‘MCDData.csv’ — Metered set of values from transformers including charge curtailment thresholds.
- ‘My Electric Avenue Published Technical Data Overview.pdf’ — Contains descriptions of data, FAQs and references to the rest of the project.
- ‘Participants.csv’ — The file containing metadata on trialists, including cluster membership.
- ‘SwitchStates.csv’ — Record of charge curtailment for the intelligent control box (ICB).

This data here, like SMiL, has a star schema, with ‘Participants.csv’ being at the center. The files contents, the variables, are tidy except for the participant’s connected phase which takes on a ‘null’ value in the case of social trialists. Also, phase information relates to the observation of a participant and cluster membership. To be tidy cluster membership must occur as a separate observation. The splitting of observations is beneficial as by maintaining distinct lists of trialists and clusters, there is a list to verify against. Further, issues emerge relating to clusters, in the participant ID field, which contains values such as: ‘ST0093’. This is a cluster code ‘ST’ followed by a non-unique membership number ‘0093’ meaning two variables are in one variable/field. Overall the rest of the dataset is tidy as several distinct series are available each recording different sets of observations.

3) *Customer Led Network Revolution*: CLNR explored several research questions gathering data from smart-meters, appliances and monitored particular assets such photovoltaic panels, electric vehicles and heat pumps. CLNR gathers it’s sub-trials into ‘test cells’, each then may be split into trialists types. Test cell labels i.e. ‘TC1a’ and ‘TC1b’ refer to the same test cell but different participants, households and SMEs respectively in this the case. Therefore each test cell contains variables possibly unique to that cell and possibly contains a subset of participants. Test cells are individually available for

⁸<https://nilm-metadata.readthedocs.io/en/latest/tutorial.html>

⁹<https://kaggle.com/jeanmidev/smart-meters-in-london>

download from the project website referenced previously, each includes a note specific to the cell.

Location IDs, equivalent to participant IDs in other trials, are again at the center of a star schema (we will continue calling them participants). Participants are linked with observations on questionnaire feedback, test cell membership as well as any metered data e.g. time series. The metered data is the same for all trials with a set of variables: $\{participantID, deviceID, parameter, value\}$ constituting the observations. This $\{parameter, value\}$ pattern or ‘tall’ style is referred to as ‘molten’ (see Wickham as cited). The entities in the schema however indicate device and measurements have their own set of observations. The files containing these each observation fully qualifies itself back to it’s participant, at the cost of repeating a lot of text, when decompressed exploding from 2 GB compressed to 60 GB! In TC1 this is entirely redundant and therefore untidy, every participant provides exactly one meter which measures exactly one parameter: the half-hourly demand. In TC2, ‘Enhanced Profiling ...’ there are multiple meters per location, however same type of units are recorded average power in kW. TC2’s messiness comes from usage of variables to store entities. Without an appliance observation, consistency is not enforced. An appliance class such as ‘fridge 1’ is the same as ‘fridge 2’ but there is no expression of this in the schema because the variable ‘meter name’ is multiple variables in one, both the ‘ID’ and the ‘class’.

CLNR is unique so far in having a schema *and* heterogeneous data. The schema reflects this with its parameter-value observations, which make it agnostic of the actual data. The implementation ignores the structuring of the data, relegating the schema to documentation. However CLNR’s ‘agnostic’ schema does not inform how sub-trials are related. Whereas the implementation acknowledges this by splitting the test cells in the repository. It could be argued that the practicality introduced by tidiness is being edged towards here.

While CLNR presents a schema suggestive of tidiness, its implementation is unwieldy and does not help to document the dataset it describes. Further issues with untidy data are in the smart meter data test cell, TC1. The two files of smart-meter data have not been merged, a note indicates they have differing observation of the UK’s daylight savings time, which is further complicated by absence of ISO 8601 dates.

IV. APPLYING TIDY DATA IN ENERGY DATA

This section applies the findings from the datasets in the previous section. The first subsection considers how a metadata schema gives a basis to understand a dataset’s structure. Next options are discussed for the OpenLV project’s data which, at the time of writing is in its closedown stages.

A. Tidiness in Datasets

The NILMtk identifies and rebuilds datasets by collaboratively implementing both a meta-schema and by writing ETL functions for each. The NILM data is typically from multiple buildings, each with a differing sets of appliances and circuit

topologies. The schema reflects this implementing a hierarchy of circuits and sub-circuits. The pattern from the NILMtk is this: start with modeling physical entities of the trial. They also maintain a global documentation of a hierarchy of appliances classes. NILM’s schema unifies the view of the trials, the implementation of the toolkit makes the dataset tidy. There is immediately complete knowledge of where and how to look for data within the dataset and a suitable assumption about its treatment and corrections. Group collaboration on schemas improves peer-review and therefore reduces errors from ETL; it is good practice of ‘Don’t Repeat Yourself (DRY)’. We argue that consensus on a data model is essential.

CLNR’s data was split into differing trials (test cells), it’s schema was valid but overlooked the complexity of realizing a dataset conforming to it. It would have benefited from a tool such as the NILMtk to manipulate it. A criticism of it’s implementation is as it used a one size fits all approach documentation about the trial was not included in a schema. This made it impossible to reuse data models and libraries of others.

We argue, generally, given a consensus on data models and convention of data the post-trial analysis will benefit greatly. Through collaboration on software and post-trial errata this provides a benefit to all using the datasets. This allows inter-dataset cross validation and application of domain best practice. Out of the box application of analysis, comparison of algorithms are all additional benefits. Future trials may model their data into the schema (or extend it appropriately), they can access similar benefits in their post-trial analysis stages and increase their post trial impact and decrease the cost of doing so.

In further application of DRY, a trial-data toolkit would reuse the pre-existing ontologies to support users in implementing and having consistent language between trials. The NILM community, developed (iteratively) their own meta-schema which has a section for entities wherein the library prescribes a language. Identifying standard lists or languages is part of the cataloguing being outlined by the UK EDTF cited in Section II.

Fig. 1: Abstracted Meta-Schema

In Fig. 1 we outline a model which can be used to define trials and datasets and encompass their entities and observations. While dedicated data repositories, such as the one SMiL resides in, provide search indexing this schema is intended to provide a basis to construct higher order analysis by permitting interrogation of semantic features the datasets. It’s constituents are:

- 1) *Meta entities* provide definitions, ideally from a common definition.
- 2) *Physical entities* provide trial specific instances from a model including serial numbers and parameters.

- 3) *Trial* provides the joining which defines parameters specific to the data collected under certain physical aspects of the trials.
- 4) *Observation* provides a domain to model the collected data by linking.

This defines how datasets are both alike in terms of meaning and alike in collection methodology. If datasets which encompass network entities both use transformers, then they will be comparable by using the same physical entities. This, like the NILMTK's schema is not a physical or logical definition but lends itself to building common software and physical implementations.

B. Example: OpenLV Project

The OpenLV project¹⁰ took place over the period 2017-2020 in the UK. A hardware and software system was installed to approximately eighty low-voltage transformers in Western Power Distribution's operational areas. Objectives were centered around the usage of this hardware to provide control, edge-processing and telemetry. This paper is concerned with the offline data, which at the time of writing, is awaiting publication as open data. The data constitutes thermal and electrical measurements at the sites taken from a number of feeders as well as ambient, enclosure and top-oil temperatures.

This project differs from the large trials; it does not have participants and therefore no star-schema. It does however have sub-trials, many of the participants have generated data which they will wish to link to aspects of the network. This dataset has three trivial physical entities: site, feeder and phase. However a suitable language to define these requires selection, the CIM is perhaps appropriate. Already two previous datasets have fields relating to these entities. This trial's electrical data for example does not necessarily relate to the transformer site itself, instead relating to feeders and phases. Permitting this data to be accessed along side other feeder data is desirable. The observations consist of various measurements taken together e.g. phase angles and complex currents. Linking physical entities to observations is especially beneficial, trials taking place attempting to control network parameters may render sites unusable for certain types of analysis. The impact, we argue, of a tidy and documented OpenLV dataset is acute, as a dataset many will use for analysis based on its snapshot of UK low-voltage networks.

V. CONCLUSION AND FUTURE WORKS

This paper suggests tidy principles are being overlooked for energy data analysis, we argue this is an acute problem. We identifying instances where tidiness is in use out of research necessity and where its absence is creating barriers and complexity to usage of datasets. Suggestions are made on how to manage tidiness for the OpenLV project's dataset. Future work is suggested in identifying suitable languages for both domains of physical entities and parameters collected.

Another direction is curation of meta schemas and collaborative work to improve library accessibility and permit building higher order analysis.

REFERENCES

- [1] H. Wickham, "Tidy Data," *Journal of Statistical Software*, vol. 59, no. 1, pp. 1–23, Sep. 12, 2014, ISSN: 1548-7660. DOI: 10.18637/jss.v059.i10.
- [2] R. Kimball and M. Ross, "ETL Subsystems and Techniques," in *The Data Warehouse Toolkit, 3rd Edition*, Wiley.
- [3] Laura Sandys, "Energy Data Taskforce report," UK, Jul. 2019.
- [4] M. Pazhoohesh, Z. Pourmirza, and S. Walker, "A Comparison of Methods for Missing Data Treatment in Building Sensor Data," presented at the 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), Aug. 2019, pp. 255–259. DOI: 10.1109/SEGE.2019.8859963.
- [5] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of Smart Meter Data Analytics: Applications, Methodologies, and Challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125–3148, May 2019, ISSN: 1949-3061. DOI: 10.1109/TSG.2018.2818167.
- [6] N. Batra, R. Kukunuri, A. Pandey, R. Malakar, R. Kumar, O. Krystalakos, M. Zhong, P. Meira, and O. Parson, "Towards reproducible state-of-the-art energy disaggregation," in *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys '19, New York, NY, USA: Association for Computing Machinery, Nov. 13, 2019, pp. 193–202, ISBN: 978-1-4503-7005-9. DOI: 10.1145/3360322.3360844.
- [7] C. Klemenjak, S. Makonin, and W. Elmenreich, "Towards Comparability in Non-Intrusive Load Monitoring: On Data and Performance Evaluation," presented at the 2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, USA: IEEE, Feb. 2020, pp. 1–5, ISBN: 978-1-72813-103-0. DOI: 10.1109/ISGT45199.2020.9087706.
- [8] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, Dec. 1992. DOI: 10.1109/5.192069.
- [9] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers, A. Singh, and M. Srivastava, "NILMTK: An open source toolkit for non-intrusive load monitoring," in *Proceedings of the 5th International Conference on Future Energy Systems*, ser. E-Energy '14, Cambridge, United Kingdom: Association for Computing Machinery, Jun. 11, 2014, pp. 265–276, ISBN: 978-1-4503-2819-7. DOI: 10.1145/2602044.2602051.
- [10] J. Kelly and W. Knottenbelt, "Metadata for Energy Disaggregation," presented at the 2014 IEEE 38th International Computer Software and Applications Conference Workshops, Jul. 2014, pp. 578–583. DOI: 10.1109/COMPSACW.2014.97.

¹⁰<https://openlv.net/>