eSim 2018
«BUILDING SIMULATION TO SUPPORT BUILDING SUSTAINABILITY»
IBPSA Canada
The Canadian Chapter of the International Building Performance Simulation Association
Le chapitre canadien de International Building Performance Simulation Association

An open data science approach for building performance studies using refitXML and Jupyter Notebooks

Steven K Firth[1], Gareth Cole[2], Tom Kane[1], Farid Fouchal[1], Tarek M Hassan[1]

[1] School of Architecture, Building and Civil Engineering, Loughborough University, UK
[2] Loughborough University Library, Loughborough University, UK

Abstract: *Open data science is increasingly viewed as a priority for collaboration and advancement of scientific disciplines. This paper presents two open data science methods for building performance studies: i) a method for storing building performance information using an open data format, including the creation of a custom XML schema and the novel use of both xml and csv formats for data storage. ii) a method for carrying out building performance analysis in an open and reproducible manner using the Jupyter Notebook tool. A Notebook is developed to demonstrate the potential of this technique for building performance modelling and analysis. The work is based on the open-access REFIT Smart Home Dataset (DOI: 10.17028/rd.lboro.2070091.v1), a published dataset of building performance information in 20 UK homes. The dataset includes detailed building survey information and over 1.3 billion sensor readings.*

Keywords: *Building performance, Smart Homes, Open data science, XML, Jupyter Notebook*

INTRODUCTION

Kraker et al. (2011) in their paper 'The case for an open science in technology enhanced learning' propose a vision for open science based on four instruments:

1. Open Access - publishing the results of academic research as freely available on the public internet
2. Open Data - publishing the datasets collected in the research process, without restricting their use
3. Open Source - making software developed in the research available under an open license
4. Open Methodology - sharing the methodology of a study, and the tools used for data collection and analysis

This could lead to many benefits for a scientific community if its members employed these instruments as part of their day-to-day research. For researchers working actively in a field, having access to previous datasets and the methods used to analyse those datasets would allow new research to develop more quickly and in a more robust fashion. There might be no need to collect new primary data if previous datasets were made available in an easy-to-use format. Data analysis techniques could be developed from previous work, reducing potential errors and the time required. Once results are generated these could be more easily checked with previous studies, improving the robustness of the results and validation of the research methodology.

Perhaps most importantly of all, active researchers who are using previous research developed with these open science instruments will have a natural blueprint for how to work in an open science manner. This makes it much easier for the current research to also employ these open science instruments and for the new research to become an open, accessible and valuable addition to the body of knowledge.

Can such a vision be a reality for the Building Simulation community? Open Access publications occur regularly in the field, with many journal papers being released in open-access form. Open Source software has been successfully demonstrated through the success of EnergyPlus which is available to use under an open license and also releases the original source code for advanced users (EnergyPlus, 2018). However Open Data and Open Methodology approaches are less well developed, with many studies not releasing either the data collected or the methods used to collect and analyse the data. This is in part because this is a difficult task to do and requires forethought and planning through the research process. It may also be because releasing the underlying data and methods of research is not generally required by funding bodies at present or generally expected by the Building Simulation community.

This paper studies the Open Data and Open Methodology instruments and their application to building performance studies. Here building performance studies refer to studies of existing buildings where data collection is carried out and data analysis is performed to generate insights into

ISBN 978-2-921145-88-6

building performance. Data collection includes both site surveys and questionnaires (to capture the characteristics of the building and occupants) and time-interval measurements made by sensors (to record dynamic performance variables). Data analysis provides insights into building performance including energy demand, internal temperatures achieved, thermal comfort of occupants, overheating, evaluation against performance metrics, evaluation against building compliance criteria and many other possible results.

Using a case study example of 20 UK homes, an Open Data and Open Methodology approach specifically suited to building performance studies are proposed in this paper. The 20 home sample was monitored intensively for an 18 month period, with over 100 sensors per home, and a detailed site survey was carried out on each building. This paper describes the methodology for preparing this detailed dataset for open-access publication and the methodology for analysing the dataset in a clear and reproducible manner. The findings illustrate the possibilities for open data science methods in building performance studies and the paper concludes with a discussion on the future of these techniques within the Building Performance and Simulation community.

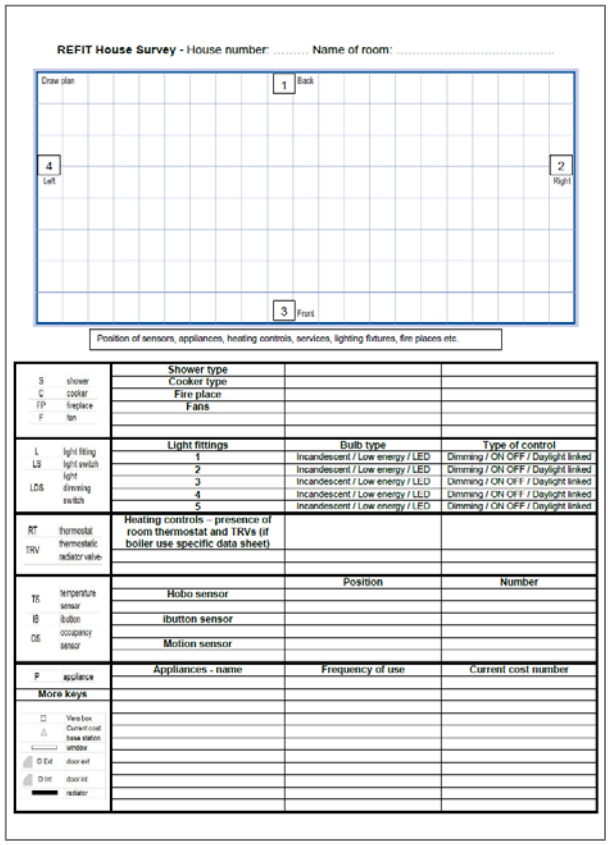
DATA COLLECTION

Overview of the REFIT study

In 2012, 20 homes were recruited in the East Midlands area of the UK to take part in a building performance study. This was part of a wider research project interested in the Smart Home concept and energy savings. The project was a UK Research Council funded project named REFIT and ran from 2012 to 2015 (REFIT, 2017). The data collected from the 20 homes forms the REFIT Smart Home Dataset which has been published as an Open Data dataset (to access the dataset follow the link in Firth et al., 2017). The process of developing this Open Data dataset is described in this section as a case study for sharing and publishing building performance data.

Original Data Collection Methods

Information about the characteristics of the homes and the people in the homes was collected through site visits. The intention was to collect sufficient data to support both the interpretation of any sensor data collected and the development of 3-D dynamic thermal models of the buildings. A review of previous work did not find any previous data collection instruments suitable for both these tasks, so a custom building survey questionnaire was developed. This enables the collection of information on the individual rooms and information about the household and the building energy systems. Figure 1 shows a page from this survey which details the information collected for each room, including the room dimensions and the presence of showers, cookers, lights, appliances, radiators, heating controls, sensors and other relevant features.


REFIT House Survey - House number: ..... Name of room: .....

Draw plan

1 Back

4 Left

2 Right

3 Front

Position of sensors, appliances, heating controls, services, lighting fixtures, fire places etc.

S shower, C cooker, FP fireplace, F fan	Shower type		
	Cooker type		
	Fire place		
	Fans		
L light fitting, LS light switch, LDS light dimming switch	Light fittings	Bulb type	Type of control
	1	Incandescent / Low energy / LED	Dimming / ON OFF / Daylight linked
	2	Incandescent / Low energy / LED	Dimming / ON OFF / Daylight linked
	3	Incandescent / Low energy / LED	Dimming / ON OFF / Daylight linked
	4	Incandescent / Low energy / LED	Dimming / ON OFF / Daylight linked
	5	Incandescent / Low energy / LED	Dimming / ON OFF / Daylight linked
RT thermostat, TRV thermostatic radiator valve	Heating controls – presence of room thermostat and TRVs (if boiler use specific data sheet)		
TS temperature sensor, IB ibutton, OS occupancy sensor		Position	Number
	Hobo sensor		
	ibutton sensor		
	Motion sensor		
P appliance	Appliances - name	Frequency of use	Current cost number
	More keys		

Figure 2: A selection of the monitoring equipment used in the project (left, Hobo pendant, Hobo U12 and iButton sensors; right, a pulse logger on a gas meter).

DEVELOPING AN OPEN DATA APPROACH FOR THE REFIT DATASET

Choice of structure

Once collected, the raw data from the study existed in a number of formats:

- Paper copies of the building survey questionnaires
- Photographs taken during and after the installation visits
- Additional notes on changes that occurred to the buildings or households throughout the study
- Notes on when and where sensors were placed in the buildings, and any changes that occurred over the study.
- csv (comma-separated variable) files holding time-series data downloaded directly from sensors
- csv files provided by external parties (for the gas meter sensor measurements which were collected remotely)

Whilst it would be possible to simply release this raw data as open access, this would not be considered good practice for an Open Data approach. A number of criteria were developed when considering the publication of this dataset:

1. No confidential data should be released
2. The data should be computer readable
3. The data should be structured for ease-of-analysis
4. The data should be fully documented, and this documentation should also be computer readable

In its raw form, the dataset did not meet these criteria. It contained confidential data, in the form of photographs and sensitive information from the building surveys, which needed removing to protect the participants' identification. Releasing scanned copies of paper forms would require each individual user to type up the data. Different interpretations might be made of the 3D geometry sketches. Many of the sensor data time periods overlapped and contained errors. This would make the dataset very difficult for users to make use of and similar analysis could lead to differing results.

A number of existing data formats were surveyed as possible candidates for the dataset publication. However none were found to be suitable in this case. BIM-based software such as REVIT could contain some, but not all, of the information recorded in the surveys. gbXML could contain the 3D geometry and building survey information, but not the more contextual information about the buildings (i.e. built form, building age) and the significant amount of sensor data recorded (gbXML, 2018). Similarly EnergyPlus idf files could be used to hold a representation of the buildings for modelling and simulation, but not a complete set of the building survey information and the sensor data (EnergyPlus, 2018).

For these reasons, a new dataset structure was developed which is named refitXML in this work. This builds on the gbxml file format which represents buildings as xml nodes, attributes and text in a tree-like structure. The purpose was not to develop a new 'standard' xml structure which would be used by other projects, but rather to create a unique data structure that was suitable for the REFIT data and one which could be adapted by other studies as needed. It is likely that a different study would ask different questions and collect information which does not fit the refitXML structure, but refitXML could be extended and adapted to suit the new study as needed.

Structure of the Final Dataset

The final REFIT Smart Home dataset has been published as an open-access dataset (Firth et al., 2017) and is stored on the Loughborough University Data Repository (Loughborough University, 2017). Figure 3 shows a screenshot of the Loughborough University Data Repository page that hosts the dataset. At the time of writing, the dataset had been cited in one journal publication (Kane et al., 2017), had been downloaded 68 times and viewed 610 times.

The REFIT project also collected additional datasets which have been made available by the partner universities. The University of Strathclyde collected over 1.5 billion power measurements from the electricity meters and appliances in the homes (available at Murray et al., 2016 and described in Murray et al., 2017). Python code used in the electricity power monitoring has been made openly available on GitHub (Murray, 2016). The

ISBN 978-2-921145-88-6

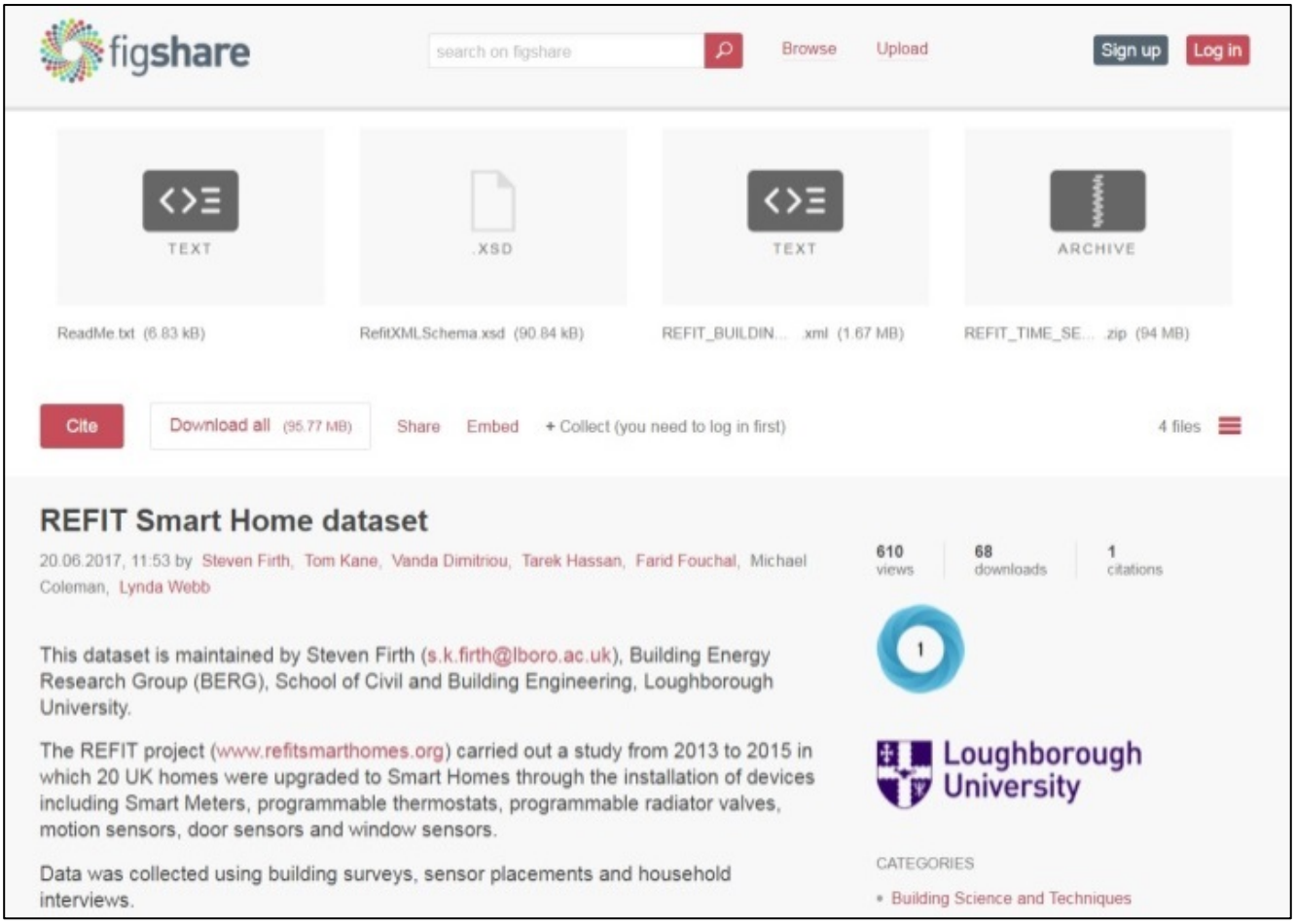


Figure 3: Screenshot of the REFIT Smart Home Dataset as hosted on the Loughborough University Data Repository (taken on 21-12-2017) (Source: Firth et al., 2017)

University of East Anglia has released survey and interview datasets from the 20 homes (Wilson et al., 2016a; Wilson el al., 2016b). The final REFIT Smart Home Dataset contained four files:

1. 'REFIT_BUILDING_SURVEY.xml'
2. 'REFIT_TIME_SERIES_VALUES.zip'
3. 'readMe.txt'
4. 'RefitXMLSchema.xsd'

The building performance data is stored in the files 'REFIT_BUILDING_SURVEY.xml', an xml file, and 'REFIT_TIME_SERIES_VALUES.zip', a zip file which contains a large comma separated variable (.csv) file. The REFIT_BUILDING_SURVEY.xml file contains the information captured through a series of building surveys, household questionnaires, site visits, sensor placements and updates about the 20 buildings throughout the study. The REFIT_TIME_SERIES_VALUES.zip file contains 25,312,397 measurements made by 2,457 sensors and other recording devices placed in the 20 homes over the 2-year study period. The dataset also contains two supporting files: the 'readMe.txt' file which provides users with basic contextual information about the dataset and how to use it; and the 'RefitXMLSchema.xsd' file, an xml schema file which is used to describe and define the structure of the REFIT_BUILDING_SURVEY.xml file.

RefitXML Schema (XSD)

The building survey data collected in the REFIT Smart Home Dataset is stored as an xml file. Xml files are ordered and structured text files, which allows for their use in a wide range of data transfer and storage processes in many applications. However for a specific application the structure of an xml file may need be specified to ensure consistency and ease-of-use by external programmes. This done using an xml schema file (.xsd) which defines and documents the structure of an associated xml file. The refitXML schema is defined by the xml schema file 'RefitXMLSchema.xsd' and is used to structure and document the building survey, interview and

ISBN 978-2-921145-88-6

sensor placement data (in fact everything except the time series data).

The first few lines of an excerpt of the schema for a Building node are shown in Figure 4. This provides a documentation string for the building node informing the user as to what the node represents.

```
<xs:complexType name="Building">
   <xs:annotation>
      <xs:documentation>
      This is a Building object.
         Buildings represent an instance
         of a building.
      </xs:documentation>
   </xs:annotation>
   <xs:choice minOccurs="0"
      maxOccurs="unbounded">
   <xs:element name="Household"
      type="refit:Household"/>
   <xs:element name="Appliance"
      type="refit:Appliance"/>
   <xs:element name="Cooker"
      type="refit:Cooker"/>
   <xs:element name="WaterOutlet"
      type="refit:WaterOutlet"/>
   <xs:element name="HotWaterCylinder"
      type="refit:HotWaterCylinder"/>
   ...
```

Figure 4: An excerpt of the 'RefitXMLSchema.xsd file which describes the structure for a 'Building' element

Building Survey Data (XML)

Whilst the schema file describes the structure of a refitXML file, the data of the REFIT Smart Home Dataset itself is stored in an instance of a refitXML file. The 'REFIT_BUILDING_SURVEY.xml' file contains this data and an excerpt for one of the house, 'Building01, is shown in Figure 5. Reading directly from the xml file, it can be seen that Building01 is a detached home, facing at 327° from North and has a cavity wall construction. Building01 contains a Space (e.g. a room) which is a heated study with a floor area of 6.25m^2. Within the study room a Hobo pendent sensor ('Sensor41') was placed which recorded the air temperature ('TimeSeriesVariable41') at 15 minute intervals from 2nd October 2013 to 3rd December 2013. This nesting of objects, sensors and variables proves to be very useful in working with the dataset.

Sensor Measurements Data (CSV)

The REFIT Smart Home Dataset contains measurement data from sensors placed in the home. This information could be stored in the refitXML file but the size of the measurement data makes this impractical. Very large xml files can be difficult to parse by external readers, and the use of tags in the xml format results in a larger file size than other formats. Instead the sensor data is stored in the standard way using a comma-separated variable (csv) file. An excerpt of the 'REFIT_TIME_SERIES_VALUES.csv' is shown in Table 1. The 'TimeSeriesVariable/@id' column is an identifier column and associated the sensor readings with a TimeSeriesVariable node in the xml file. The 'dateTime' column provides a timestamp for the sensor readings using the standard xml date and time format. The 'data' column shows the value of the sensor reading itself. In this way, Table 1 shows that on the 2nd of October 2013 at 05:00 variable TimeSeriesVariable1 had a value of 17.772. To understand what this value represents, the user must look up the information about TimeSeriesVariable1 in the separate xml file (in this case the 17.772 value is an air temperature reading by a Hobo U12 sensor recorded in a bathroom in Building01).

Table 1: An excerpt of the REFIT_TIME_SERIES_VALUES.csv file showing the first 10 rows of data

TimeSeriesVariable/@id	dateTime	Data
TimeSeriesVariable1	2013-10-02T05:00:00Z	17.772
TimeSeriesVariable1	2013-10-02T05:30:00Z	18.081
TimeSeriesVariable1	2013-10-02T06:00:00Z	18.176
TimeSeriesVariable1	2013-10-02T06:30:00Z	18.176
TimeSeriesVariable1	2013-10-02T07:00:00Z	18.105
TimeSeriesVariable1	2013-10-02T07:30:00Z	18.01
TimeSeriesVariable1	2013-10-02T08:00:00Z	17.891
TimeSeriesVariable1	2013-10-02T08:30:00Z	17.772
TimeSeriesVariable1	2013-10-02T09:00:00Z	17.701
TimeSeriesVariable1	2013-10-02T09:30:00Z	17.677
…	…	…

ISBN 978-2-921145-88-6

```
<Building id="Building01" startDateTime="2013-10-01T00:00:00Z" occupancyType="Single
   family dwelling" builtFormType="Detached house or bungalow" orientation="327"
   wallTypeMainBuilding="Masonry-Boxwall-Cavity" wallAgeBandMainBuilding="1975 - 1980"
   cavityWallInsulationPresent="Yes" windowType="Double glazed - UPVC" loftType="Fully
   boarded" loftInsulationType="Mineral wool/fibre glass" loftInsulationThickness="300mm">
   ...
   <Space id="Space1" startDateTime="2013-10-01T00:00:00Z" conditionType="Heated"
      area="6.25" volume="14.375" storeyLevel="0" roomType="Study">
      <Sensor id="Sensor41" startDateTime="2013-10-02T05:00:00Z" endDateTime="2013-12-
         03T15:15:00Z" manufacturer="Onset" model="Hobo pendant">
         <TimeSeriesVariable id="TimeSeriesVariable41" startDateTime="2013-10-02T05:00:00Z"
            endDateTime="2013-12-03T15:15:00Z" variableType="Air temperature" units="C"
            intervalType="FixedInterval" intervalUnit="Minute" intervalLength="15"
            hasMissingData="No" repeatsOmitted="No" hasDuplicateTimestamps="No"/>
      </Sensor>
      ...
   </Space>
   ...
</Building>
```

Figure 5: An excerpt of the REFIT_BUILDING_SURVEY.xml file which describes an actual building 'Building01' and the properties of and relationships between Building01, Space1, Sensor41 and TimeSeriesVariable41.

DEVELOPING AN OPEN METHODOLOGY FOR DATA ANALYSIS

Choice of Open Methodology approach

Data analysis is an essential part of generating meaningful results and insights from the data collected in Building Performance Studies. In making data analysis open, reproducible and accessible, the following criteria were developed:

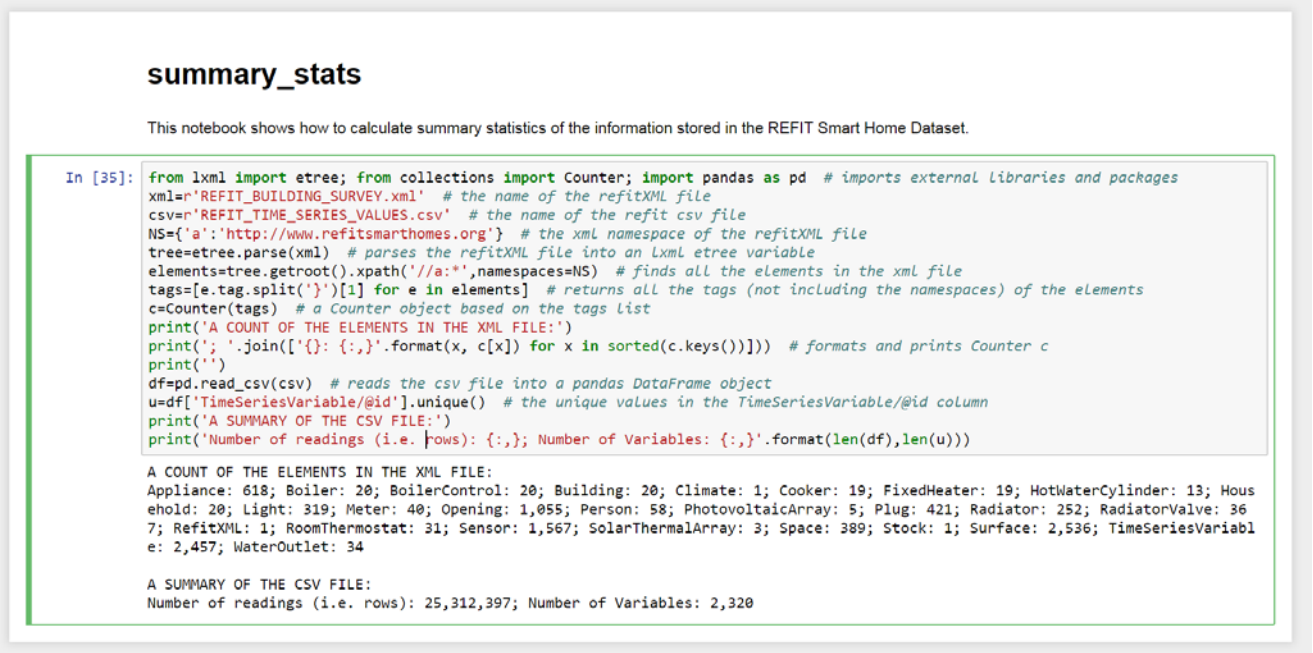
1. Data analysis methods should work directly on a dataset
2. All data cleaning, processing and analysis steps should be fully documented and reproducible (in practice the only way to fully document these steps is using computer code)
3. No manual manipulation of the dataset should take place (as this is implicit and cannot be fully documented)
4. Computer code should use the best practice in software design including variable naming conventions, line comments, documentation strings and code-reuse (through functions and classes).

Using software such as Microsoft Excel is unlikely to meet these criteria as the simple act of copying and pasting a selection of data will results in the data analysis steps not being fully reproducible. Through a review of data analysis software, in this work the choice was made to develop the analysis code in the Python programming language, as Python was specifically designed to be easy-to-read and share (Python, 2018). The Python analysis code is shared within Jupyter notebooks, which can hold the code itself, the outputs of the code and additional written documentation (Project Jupyter, 2018). Jupyter notebooks provide a good solution to enabling other researchers to reuse the data analysis methods, and users of a notebook can experiment with the results by altering the code within a Notebook and rerunning the results. The documentation describes Jupyter Notebooks as:

> *"Notebook documents (or "notebooks", all lower case) are documents produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc...). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc..) as well as executable documents which can be run to perform data analysis." (Jupyter Guide, 2018)*

Jupyter notebooks are not a new programming language but rather a way of holding both programming code and documentation about the code in the same place. This suits the Open Methodology approach, which aims not simply to provide code to run but also full documentation of the methods used and enable other researchers to reproduce the results.

ISBN 978-2-921145-88-6



summary_stats

This notebook shows how to calculate summary statistics of the information stored in the REFIT Smart Home Dataset.

In [35]:
```
from lxml import etree; from collections import Counter; import pandas as pd  # imports external libraries and packages
xml=r'REFIT_BUILDING_SURVEY.xml'  # the name of the refitXML file
csv=r'REFIT_TIME_SERIES_VALUES.csv'  # the name of the refit csv file
NS={'a':'http://www.refitsmarthomes.org'}  # the xml namespace of the refitXML file
tree=etree.parse(xml)  # parses the refitXML file into an lxml etree variable
elements=tree.getroot().xpath('//a:*',namespaces=NS)  # finds all the elements in the xml file
tags=[e.tag.split('}')[1] for e in elements]  # returns all the tags (not including the namespaces) of the elements
c=Counter(tags)  # a Counter object based on the tags list
print('A COUNT OF THE ELEMENTS IN THE XML FILE:')
print('; '.join(['{}: {:,}'.format(x, c[x]) for x in sorted(c.keys())]))  # formats and prints Counter c
print('')
df=pd.read_csv(csv)  # reads the csv file into a pandas DataFrame object
u=df['TimeSeriesVariable/@id'].unique()  # the unique values in the TimeSeriesVariable/@id column
print('A SUMMARY OF THE CSV FILE:')
print('Number of readings (i.e. rows): {:,}; Number of Variables: {:,}'.format(len(df),len(u)))
```

```
A COUNT OF THE ELEMENTS IN THE XML FILE:
Appliance: 618; Boiler: 20; BoilerControl: 20; Building: 20; Climate: 1; Cooker: 19; FixedHeater: 19; HotWaterCylinder: 13; Hous
ehold: 20; Light: 319; Meter: 40; Opening: 1,055; Person: 58; PhotovoltaicArray: 5; Plug: 421; Radiator: 252; RadiatorValve: 36
7; RefitXML: 1; RoomThermostat: 31; Sensor: 1,567; SolarThermalArray: 3; Space: 389; Stock: 1; Surface: 2,536; TimeSeriesVariabl
e: 2,457; WaterOutlet: 34

A SUMMARY OF THE CSV FILE:
Number of readings (i.e. rows): 25,312,397; Number of Variables: 2,320
```

Figure 6: Browser screenshot of a Jupyter notebook which calculates summary statistics about the composition of the REFIT Smart Home Dataset.

Data Analysis Example: Summary statistics

Figure 6 shows a screenshot of a Jupyter notebook. Notebooks appear and are edited within an internet browser (linked to a local server) and so the figure shows a screenshot of the browser itself with the top and bottom sections cropped. The notebook shows two 'cells'. The upper cell is a text cell and in this example shows a title and basic description about the notebook. Text in these cells is written in 'markdown', a simple tag-based language which can provide formatting such as headers, bold, bullet points etc. The lower cell contains computer code, written in Python, in the grey shaded box and also the output of the code once it is run in the area below the grey shaded box.

The outputs of the notebook in Figure 6 show the reader a number of basic summary statistics about the REFIT Smart Home Dataset, such as the number of appliances recorded (618) and the number of readings recorded by the sensors (25,312,397). However, and more importantly in this paper, it shows the exact method of how these results were calculated. The Python code imports the xml and csv files of the REFIT dataset directly and carries out the analysis tasks to provide these statistics. This is a complete process, and the steps taken from the original data to the final results can be exactly traced by other users. Similar notebooks could be designed for creating publication-ready tables and plots directly from the dataset.

DISCUSSION

Limitations of the approaches presented

This paper has presented solutions for the Open Data and Open Methodology instruments in the context of Building Performance studies. The methods given here are not intended as final or generic solutions, but rather represent a first step and illustrative approaches which others may adapt for future studies.

The Open Data approach based on refitXML represents a solution which was specifically developed for the data collected in the REFIT project. It represents a significant improvement on simply releasing the raw data as collected, but a number of limitations still exist. Firstly in the published dataset itself the original data collection instruments (such as the blank paper building surveys etc.) have not been provided. This was partly intentional, as the questions themselves and the response options are documented in the refitXML schema file so the original data collection instruments may not be required. However some users may prefer to have the actual questions, and the format and structure of the questions within the questionnaire. The original building survey were designed as an in-house method and never intended for public

ISBN 978-2-921145-88-6

release, and might require further editing and documentation. It is still an open question though as to whether users require this feature.

Secondly some important information is not provided in the public version of the dataset due to confidentiality issues. In particular the 3D geometry of the buildings was not made publically available as it was deemed that this could potentially be used to directly identify the individual households. This data is retained in an internal version of the dataset but is stripped out of the public release with only the orientation and areas of surfaces made available. As such, the dataset cannot be used as it stands for 3D building simulation modelling. An important feature of future data collection will be to establish agreement at the start of new projects which will allow such information to be made public.

Thirdly the use of the xml format has proved challenging for some users. Computer scientists and programmers have found the structure logical and simple to use, but other users more familiar with the table structures of Excel or relational databases have initially struggled to interpret the dataset. It could be argued that the choice of using xml is failing the criteria that the Open Data should be structured for ease-of-use. However in this case, for a study which captures not only building level information but information on a multitude of objects within the building itself, structuring the dataset in database-style tables does not simply the structure. Rather the equivalent database contains 30+ tables with many primary-foreign key links which results in long and complex SQL query expressions. For computer scientists and others who understand the syntax, the use of xml simplifies data queries by using the XPATH query language. There is a trade-off here, using the xml structure may result in a dataset which is harder for some users to interpret but the xml structure is likely to result in simpler data analysis code.

The main limitation of the Open Methodology approach for the data analysis is the use of the Python programming language which will only be familiar to a subset of users. However if the conclusion is accepted that programming code is the only way to create truly documented and reproducible data analysis, then this is an intrinsic limitation. The choice of Python is reasonable here as one of the most popular programming languages for data analysis, and one of the easiest to learn. Another limitation may be the complexity of the code used to create the results. This appears to be largely a function of the structure of the dataset itself. A better structure of the dataset should result in simpler logic and more intuitive analysis techniques.

The Future of Open Data Science for Building Performance Studies

The solutions presented in this paper have shown that it is possible to use Open Data and Open Methodology approaches in Building Performance studies. These solutions are intended as illustrative and not final, generic approaches for all cases.

One vision for the implementation of these approaches is the publication of journal papers which both release the data used in an Open Data fashion and release the data analysis methods used as an Open Methodology. A further step could be to publish an entire journal paper as a Jupyter notebook, where the text, table and figures of the paper are provided, and the programming code to create the results is also present. This would create a single document in which the results of a paper and the methods used to create the results are combined in a single file, and enable other researchers to study and modify the data analysis methods for their own studies.

The solutions in this paper could equally be used for building simulation studies which employ modelling as the main method for results generation. Input files to models (such as EnergyPlus idf files) can be made open access and the data analysis steps used to analyse the model output files (such as EnergyPlus csv results files) could be provided as Jupyter notebooks. In a parametric analysis, the code used to modify model input files could also be made open.

The ease of reuse of data analysis techniques and programming code is largely impacted on the data structures under analysis. One could imagine a series of Python functions and classes for analysing EnergyPlus results files and, because this is a clearly defined and widely used format, these function and classes could be easily used by others. For building survey data, BIM formats such as IFC and gbxml may provide a common format which could be widely used. However these approaches would need to be further developed for widespread use, or new approaches may be required such as the refitXML format as described in this paper. Improving the data structures used within Building Performance studies and Building Simulation could be a useful goal for the Building Simulation community.

CONCLUSION

This paper has described the collection of building survey and sensor data for 20 homes in the REFIT project. The paper describes the benefits of using an open science approach for the publication of this dataset and in the analysis of the data. An Open Data approach using refitXML is described for developing an open, structured

ISBN 978-2-921145-88-6

and documented dataset from raw data which has been made publically available for other researchers to use. An Open Methodology approach is also described which creates an open, reproducible and documented method for data analysis using Python and Jupyter notebooks.

The key conclusions are:

- It is possible to create building performance datasets based on Open Data principles. However the existing data formats are not suited to all studies and new format may need to be developed, such as the refitXML format as described in this paper.
- Python and Jupyter notebooks provide a solution to publishing data analysis techniques and meeting Open Methodology requirements.
- The data structure used in building performance studies is a key factor in both the ease-of-interpretation of the dataset, and the complexity of programming code used in the analysis of the dataset.

Further work includes further development of data structures and formats for Building Performance studies and Building Simulation studies. More development of the data analysis techniques is also planned, with the goal of publishing an academic journal paper in which the data analysis methods are fully documented and published using the Jupyter notebook approach.

ACKNOWLEDGEMENT

This work has been carried out as part of the REFIT project ('Personalised Retrofit Decision Support Tools for UK Homes using Smart Home Technology', Grant Reference EP/K002457/1). REFIT is a consortium of three universities - Loughborough, Strathclyde and East Anglia - and ten industry stakeholders funded by the Engineering and Physical Sciences Research Council (EPSRC) under the Transforming Energy Demand in Buildings through Digital Innovation (BuildTEDDI) funding programme. For more information see: www.epsrc.ac.uk and www.refitsmarthomes.org.

REFERENCES

EnergyPlus. 2018. https://energyplus.net/

Firth S.K., Kane T., Dimitriou V., Hassan T.M., Fouchal F., Coleman M. and Webb L. 2017. *REFIT Smart Home dataset*, figshare, https://dx.doi.org/10.17028/rd.lboro.2070091

gbXML (2018). *gbXML: an industry supported schema for sharing building information between disparate building design software tools*, available at http://www.gbxml.org/

Kane T., Firth S.K., Hassan T.M. and Dimitriou V. 2017. *Heating behaviour in English homes: An assessment of indirect calculation methods*, Energy and Buildings, 148, pp.89-105 http://dx.doi.org/10.1016/j.enbuild.2017.04.059

Kraker P., Leony D., Reinhardt W. and Beham G. 2011. *The case for an open science in technology enhanced learning, Int. J. Technology Enhanced Learning*, Vol. 3, No. 6, pp 643-654, http://www.know-center.tugraz.at/download_extern/papers/open_science.pdf

Jupyter Guide. 2018. *Jupyter/IPython Notebook Quick Start Guide*, http://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html#notebook-document

Loughborough University. 2017. *The Loughborough University Data Repository website hosted by Figshare.com*, https://lboro.figshare.com/

Murray D. and Stankovic L. 2016. *REFIT: Electrical Load Measurements (Cleaned)* - available at: http://dx.doi.org/10.15129/9ab14b0e-19ac-4279-938f-27f643078cec

Murray, D. 2016. *Code used for the REFIT project data acquisition*, GitHub – available at: https://github.com/David-Murray/REFIT

Murray D., Stankovic L., and Stankovic V. 2017. *An electrical load measurements dataset of United Kingdom households from a two-year longitudinal study*, Scientific Data, vol. 4, Article number: 160122. http://dx.doi.org/10.1038/sdata.2016.122

Project Jupyter. 2018. http://jupyter.org/

Python. 2018. https://www.python.org/

REFIT. 2017. *The REFIT project website*, www.refitsmarthomes.org

Wilson C. and Hargreaves T. 2016. *REFIT: Personalised retrofit decision support tools for UK homes using smart home technology. Phase 1: Survey data*. [Data Collection]. Colchester, Essex: UK Data Archive. 10.5255/UKDA-SN-852366 - available at: https://dx.doi.org/10.5255/UKDA-SN-852366

Wilson C., Hargreaves T. and Hauxwell-Baldwin R. 2016. *REFIT: Personalised retrofit decision support tools for UK homes using smart home technology. Phase 2: Smart home interviews*. [Data Collection]. Colchester, Essex: UK Data Archive. 10.5255/UKDA-SN-852367 - available at: https://dx.doi.org/10.5255/UKDA-SN-852367