LBNL-6724E

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Data Preparation Process for the Buildings Performance Database

Claudine Y. Custodio, Travis Walter, Laurel Dunn,
Andrea Mercado, Richard E. Brown and Paul Mathew

Environmental Energy Technologies Division

August 2015

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California

Acknowledgements

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Program, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The authors would like to thank

- Michael D. Sohn, Lawrence Berkeley National Laboratory
- D. Magnus Cheifetz, Building Energy Inc.
- Elena Alschuler, U.S. Department of Energy
- Jessica Knapstein, Energetics Inc.
- Benjamin Laurent, National Renewable Energy Laboratory
- Micah Brill, Urban Land Institute

Abstract

The Buildings Performance Database (BPD) includes empirically measured data from a variety of data sources with varying degrees of data quality and data availability. The purpose of the data preparation process is to maintain data quality within the database and to ensure that all database entries have sufficient data for meaningful analysis and for the database API. Data preparation is a systematic process of mapping data into the Building Energy Data Exchange Specification (BEDES), cleansing data using a set of criteria and rules of thumb, and deriving values such as energy totals and dominant asset types. The data preparation process takes the most amount of effort and time therefore most of the cleansing process has been automated. The process also needs to adapt as more data is contributed to the BPD and as building technologies evolve over time. The data preparation process is an essential step between data contributed by providers and data published to the public in the BPD.

TABLE OF CONTENTS

Introduction

The Buildings Performance Database, available at the http://bpd.lbl.gov, allows users to apply statistical analysis on a dataset of anonymous empirical data for commercial and residential buildings. The BPD allows users to select a peer group of buildings based on various database filters including building type, location, size, vintage, equipment and operational characteristics. The users can also change database filters to evaluate differences in the distribution of energy use intensities for peer groups with different equipment types or operational differences. In addition, statistical analysis provided in the BPD allows users to see the potential energy use difference of having one building system over another. The BPD is powered by a database containing the characteristics and energy consumption data from buildings across the United States. For the statistical analysis to give justifiable results, data preparation is an essential process between data provided to the BPD team and data uploaded into the database.

The BPD includes empirically measured data—not modeled data or from anecdotal evidence—from a variety of data sources such as building owners, property management companies, energy-efficiency programs, and utilities. Given the widely varying purposes for data collection, as well as differences in collection methodology, data contributed to the BPD are in many different formats and with a range of data quality. For example, some datasets provide granular information about building characteristics (e.g., equipment capacity or efficiency, operation schedule, control systems, etc.), while others provide only the barest information about floor area and building use. Likewise, energy consumption data can be provided over a range of sampling intervals, from annual or monthly values drawn from utility bills, to hourly or 15-minute values from interval meters.

Data preparation is the process of systematically reviewing data, and removing suspected erroneous data based on rules of thumb and criteria. Databases populated with data that have not gone through a data preparation process will vary more in data quality, making it more difficult to compare within or between peer groups of buildings. The standard data format used in the BPD is the Building Energy Data Exchange Specification (BEDES 2015).

This report documents the standardized data preparation process the BPD team uses to provide transparency to the methods used in getting raw data from providers to the data displayed to the public through the website. The data preparation process maps, cleans, and derives values from user-provided data before being uploaded to the BPD database. The result of the process is a dataset that can be statistically analyzed to provide meaningful results. This report includes a summary of the data preparation process and rules currently applied to datasets, lessons learned by the BPD team from preparing numerous datasets for the database, and outlines future improvements being considered.

Methods

The data preparation process for a specific dataset can vary depending on the size and format of the dataset, the number of data fields, and the quality of the raw data. Datasets submitted to the BPD go through a multi-stage process to ensure that data in the BPD database are consistent and reliable.

When the LBNL team receives data from providers the data is mapped to a common schema. The mapped data is then cleansed using rules based on building science, previous studies, and industry experience. In addition to the cleansed data, derived values are calculated. Finally, the cleansed data and derived values are uploaded to the database and can be queried by the filtering and analysis algorithms (Figure 1).

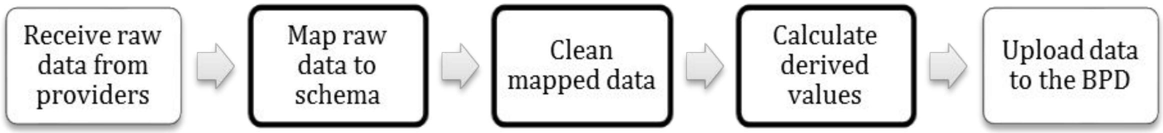


Figure 1. Buildings Performance Database process for preparing datasets for the tool.

For large portfolios of buildings spanning multiple sectors, data collection processes, or building managers, it is possible that the original data will contain erroneous, corrupt, or missing data. The data preparation process reconciles differences in data quality and availability to make the aggregated dataset more uniform and consistent, so analysis methods are more accurate and efficient. Some of these problematic data may be easy to identify, such as zero floor areas, or an energy consumption value off by several orders of magnitude. Other errors can be difficult to identify. For example, a building that appears to be ten times less efficient than its peers might be a decimal error, or can be explained by a unique energy end-use, such as a data center. While no data preparation process is flawless, the team emphasizes the importance of having a consistent and scientifically sound data preparation procedure to ensure the final database does not contain glaring errors such as disproportionate energy use intensities (EUI's) or misrepresents buildings with estimated or default values.

1. Mapping

The mapping step involves translating the raw data fields into a standard data schema shown in Appendix A: Buildings Performance Database Data Schema. The BPD uses BEDES as its standard data format. BEDES supports collection and analysis of empirical building performance data by providing a common data format, definitions and an exchange protocol. Further information about BEDES can be found at . The mapping process involves: matching raw data field names with the schema field names, translating building and asset types to schema enumerated types, converting data types, parsing text descriptions, and other processes that encodes raw data from providers into a uniform scheme. Mapping is an important first step because data need to be standardized before cleansing rules and checks are applied. Most datasets adhere to their own schema or data format. Ideally, data providers include a reference table to their data fields and enumerated values, which makes mapping easier. Often, data providers do not include reference tables and it is up to the data processor to devise their own reference table based on interviews with the data provider, other inputs in the same dataset, and familiarity with building technologies.

For instance, a raw data field is named "TYPE" or "BLDG TYPE" to designate the building type for their individual records. The equivalent BEDES schema for this field can be the “Residential Facility Type” or “Facility Type” for commercial facilities. Under “TYPE” the dataset could have “Middle School” and “High School”, which in BEDES is, “Education - Elementary/Middle School” and “Education - High School” respectively. Continuing with the example, both properties report 50% heated floor areas. To comply with allowable values, the percent of floor area is multiplied by the reported floor area to fill in the “Heated Floor Area” field. If the heating equipment is reported as “heat pump, air src” and a column named “Programmable Thermostat” has True, the equivalent in BEDES is to map “Heat Pump - Air Source” as the “Heating Type” and “Programmable Thermostat” as the “Control Type”.

The more asset information tied to each building or facility, e.g., HVAC equipment and other building characteristics, the more accurately the building can be categorized in the database and queried by the web application.

2. Cleansing

The cleansing step involves reviewing consistency within and across data fields. In this step, the data processor determines whether to include the facility in the database and delete only the erroneous data fields, or to exclude the entire facility from the database. These decisions are informed by rules developed by the BPD team based on building science and rules of thumb. These checks are more effective when the mapping process is done correctly. The most rigid of these rules is the minimum data checklist. Facilities that do not meet these requirements are considered incomplete and are excluded from the database. The minimum data for the BPD are: gross floor area and primary facility type. The next step is checking for allowable values. This process verifies that data fields are filled with the correct data type, enumerated values are translated to their BEDES equivalents, and numerical values are within the range of realistic values. The final set of rules confirms consistency of values within a facility.

The cleansing process is an iterative process. Erroneous data is removed and the cleansed data is put through the cleansing process again. For efficiency, most of the cleansing rules are automated in scripts. A full list of cleansing rules is included in Appendix C: The BPD Cleansing and Validation Rules. All of the data in the BPD were verified using the rules before they are merged with the public database.

Minimum Data Checklist

The minimum data checklist is a list of requirements for a building to be considered a complete entry. Data submitted for the BPD must have gross floor area and the primary facility type. One continuous year of energy measurements is highly preferred, but not strictly required.

The location data is used for peer group filtering and to determine the climate zone, if not already provided, which is also used for filtering peer groups in the BPD tool. The facility gross floor area and the energy data are used to calculate the energy use intensity of the facility, the main metric used in comparing and analyzing the facilities. The facility type is a required tag to understand the differences in building energy use. The major fuels used in a building are defined as the fuels necessary to operate and maintain building systems, mainly electricity and when available, natural gas. The major fuels were determined based on the commonly metered fuel streams in CBECS and RECS. For a full list of the minimum data checklist refer to Appendix D: Minimum Data Requirements. Datasets or individual properties that do not meet these requirements are not included in the BPD.

The checklist, mainly the requirements for energy records, ensures that the BPD algorithms will provide justifiable EUI results for comparing building technologies. The building's annual energy use intensity (EUI) is the main metric reported back to users and used to drive algorithms for evaluating energy performance. The EUI is a metric that is easily calculated from the minimum data requirements and is used widely in benchmarking databases such as ENERGY STAR's Portfolio Manager or national surveys such as the Commercial Buildings Energy Consumption Survey (CBECS), the Residential Energy Consumption Survey (RECS), and California End-Use Survey (CEUS). Calculated energy use intensity should be consistent with CBECS, RECS or CEUS EUI ranges.

The EUI is calculated for whole-building consumption. Assume the facility as a control volume, where all energy is consumed completely and only in that facility (a facility can be a campus, a free-standing building, or an area within or part of a building). Zero or negative EUI values for electricity use should raise a flag, but are not a sufficient condition for excluding the building. Zero values or negative EUI values can be explained by periods of vacant or by net metering. Currently, buildings with energy generated on-site, typically net-metered, are included in the BPD. Records from central plants are excluded unless the whole facility includes the central plant and all the buildings it serves.

Fuel used in large quantities and reported to serve a specific piece of equipment (e.g., gas boiler) should be present in the individual fuel streams. Having a continuous energy record is in accordance with the BPD philosophy of populating the database with real life empirical data since missing values would require interpolation and could misrepresent the building on record.

These strict rules were devised with the recommendations from the BPD analysis team in order to have a uniform and consistent format to compare, analyze, and query data from the BPD database.

Allowed Values and In-range checks

The rules governing allowable values are based on the data type and realistic values for each field, for example, percent values should be between 0 to 100. If the data field is part of the minimum data checklist then the facility is excluded from the database. For instance, a facility with a gross floor area that is negative or zero is deleted. If the data field is not part of the minimum data checklist, the only the inputs that do not meet these criteria are excluded from the database. Therefore, a facility can be missing "Average Weekly Operating Hours" if the reported hours are greater than 168, the total number of hours in a week. The main guideline is to follow BEDES for values under the specific data field. For example, "Heating Equipment" should be one of the enumerated types listed in BEDES. See Appendix C: The BPD Cleansing and Validation Rules for a full list of the allowed values and in-range checks.

Review for Consistency

After verifying that the data fields are filled in correctly, each database record is then checked for consistency throughout related data fields. For example, the sum of "Activity Floor

Area" in a facility should be less than or equal to the "Gross Floor Area" of the facility. Input values that do not meet consistency criteria are excluded from the database, but we do not necessarily exclude the whole facility. See Appendix C: The BPD Cleansing and Validation Rules for a full list of consistency checks.

Dataset Specific Rules

Even with the multiple checklists, there are cleansing decisions that are dataset specific. As mentioned earlier datasets often follow a schema different from what the BPD uses or the data quality is not as stringent. Dataset specific rules are not applicable to the majority of datasets submitted to the BPD.

In the following example, a number of interpolated values were identified after applying the standard cleansing rules. This difference in data quality standards between the BPD and one of its data contributors became apparent only through further analysis of the dataset.

The dataset consists of over 11,000 residential buildings in the Northeastern United States. About 8% of these buildings were removed during the initial cleansing due to failure to meet the cleansing thresholds prescribed in the cleansing rules. A histogram of the dataset after applying standard cleansing rules, shown in Figure 2(a), reveals that nearly 8% of the remaining buildings have EUI's equal to 32 kBtu/ft^2/year. Although data are not expected to follow smooth distributions, the unusual spike in the distribution is suspect and prompted further investigation. Correspondence with the data provider confirmed that energy use for some records were estimated using a default EUI for buildings that were unable to provide a complete year of energy data. To guarantee the quality of data imported into the BPD database from a dataset identified to contain interpolated data, buildings with estimated EUI were removed.

An additional cleansing action was devised to remove the buildings that reported estimated energy use. This issue is unique to this specific dataset. While this additional cleansing action may be considered excessively restrictive for most dataset, the BPD prioritizes data quality over number of buildings and therefore implemented a rule that would guarantee high data quality. With this additional cleansing rule, 16% of the buildings in original dataset were eliminated.

Analysis of the fully cleansed data reveals some minor differences in the overall characteristics of the dataset shown in **Table 1**. Visual comparison of Figure 2(a) and Figure 2(b), confirms that there is no longer a spike in the histogram near 32 kBtu/ft^2/year, which initially caused suspicion about the data quality. The histogram follows an approximately log-normal distribution, which is consistent with expectations based on previous studies, further establishing confidence in the quality of the remaining data.

Table 1. Dataset characteristics, before and after standard and supplemental cleansing.

Data Characteristic	**Before Cleansing**	**After Cleansing**
Number of Buildings	11,485	8,758
Aggregate Gross Floor Area	22 million ft^2	17 million ft^2
Total Annual Energy Consumption	1.4 billion kBtu	0.97 billion kBtu
Median Annual EUI	52.5 kBtu/ft^2/year	56.7 kBtu/ft^2/year

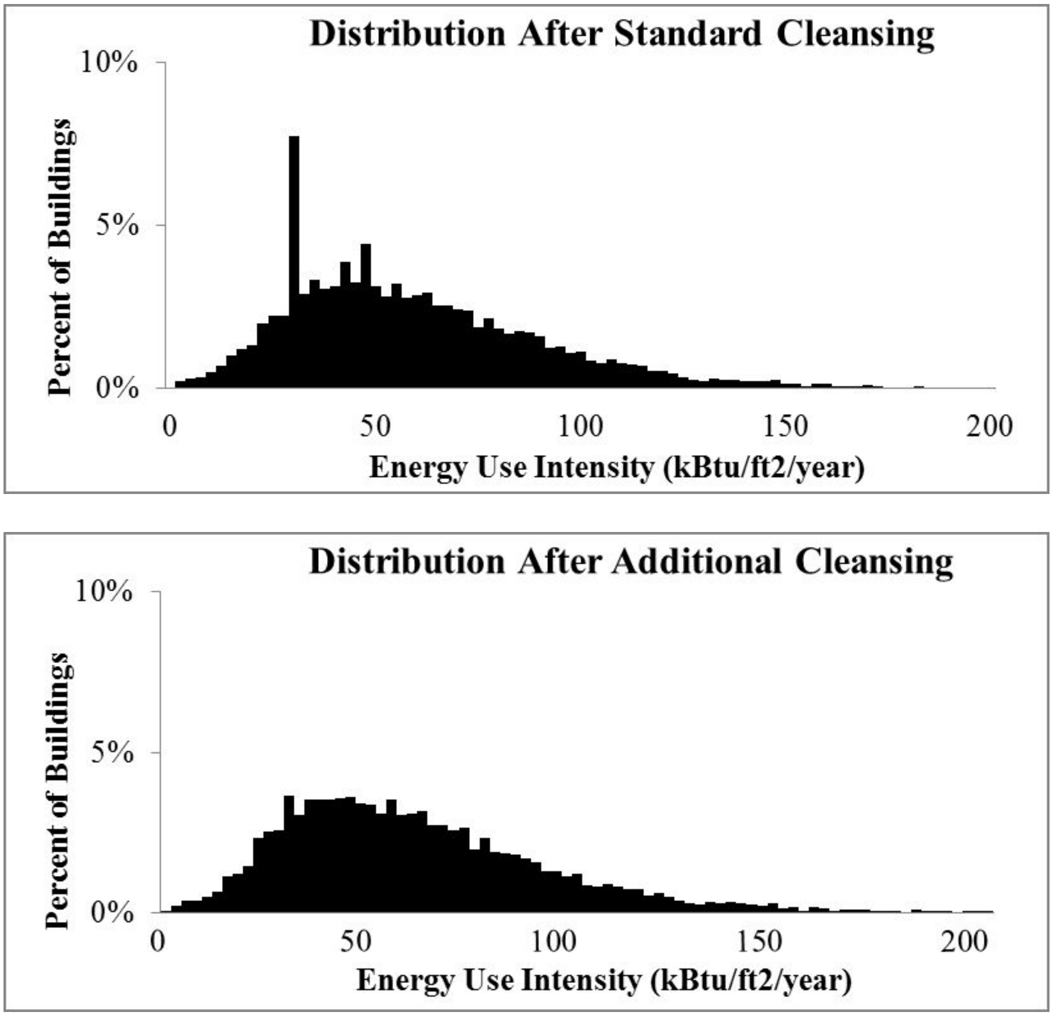


Figure 2. Histograms showing the EUI distribution of a residential dataset in the Northeastern United States (a) after standard cleansing [top] and (b) after further cleansing designed to eliminate an identified data quality issue [bottom].

3. Computing Derived Values

The primary facility type, equipment types and energy totals are values that are used in the BPD tool algorithms and filtering. Further processing of the data contributed to the database is required for algorithms to work. These derived values are determined using related data fields such as “Lighting Type” and “Lighting Capacity”.

Dominant Facility and Equipment Types

The database frequently contains buildings that fit partially in one building type and partially in another. Buildings submitted to the BPD often have multiple activity areas or multiple types of heating or cooling equipment, but the energy use is reported for the whole building. When applying filters to the database to select peer groups, and when computing predictors for the multiple linear regression model, it is necessary for a building to be classified as only one category. The dominant types represent where most of the energy use is allocated for the building. The rules for assigning dominant types are discussed in more detail below.

Facility Type

Facility Types are the use types of the building or facility. Facility types are similar to the Principal Building Activity field in CBECS whose purpose is to group buildings according to consumption patterns. A building can have multiple facility types but are connected to a single master meter or report energy use for the whole buildings.

For most residential buildings, one facility type is reported for the entire facility, so dominant type calculations are not necessary.

For commercial buildings, the gross floor area can be divided into different use cases, what is called “Activity Types” in BEDES. If at least 80% of the building's gross floor area is devoted to an activity type, that activity type is designated as the building's primary facility type. If no activity types meet this criterion, or if the data is missing, the next step is to check the number of activity area records reported for the building. If there is only one activity type, that type is designated as the primary facility type. Finally, if none of the above criteria are met, the building's primary facility type is “Commercial – Uncategorized”.

An exception to the above rule is mixed-use buildings, which contain both commercial and residential uses. For this case the following rules apply:

- If at least 80% of the building's gross floor area is devoted to residential use, the building's facility type is chosen to be “Mixed Use – Predominantly Residential”.
- If 20% or less of the building's gross floor area is devoted to residential use, the building's facility type is chosen to be “Mixed Use - Predominantly Commercial”.
- If the residential floor area is between 20% and 80% of the total gross floor area, the building's facility type is chosen to be “Mixed Use - Commercial and Residential”.

Operating Hours

The operating hours is the total number of hours in a week that the building is in use or occupied.

For residential buildings, operating hours are not specified, so calculations are not necessary.

For commercial buildings, average weekly operating hours are defined for each activity area. A building's operating hours are computed by simply averaging the operating hours of each activity area. Activity areas with empty values for operating hours are ignored.

Wall R-Value

The wall R-value is a measurement of building insulation.

Dominant wall R-values are given for each property with available wall R-value. A building's wall R-value is computed by summing the R-values of each record, weighted by the proportion of the building's wall area reported for that wall record. Wall records with empty values for wall R-value or wall area should be ignored. If there is only one wall record reported for the building, that wall R-value is chosen.

Lighting

Lighting types are the artificial sources of ambient lighting installed in a building.

If at least 80% of the building's installed ambient lighting power is of a particular lighting type, that type is chosen as the building's dominant lighting type. If the criteria is not met using the “Installed Power” data, or if the data is missing, the criteria is then tested using the “Percentage of Total Installed Power” data. If the criteria are still not met, or if the data is missing, the dominant type is based on “Percentage of Total Floor Area Served”. If at least 80% of the floor area served is of a particular lighting type, that lighting type becomes the dominant lighting type. If the criteria are still not met, or if the data is missing, the next step is to check the number of lighting records reported for the building. If there is only one lighting type, that lighting type is chosen as the dominant type. Finally, if none of the above criteria are met, the building's dominant lighting type is “Other Or Combination”. When a building does not report a Lighting Type, the default value is “Unknown”.

Flow Control System

Flow control systems are the mechanisms the building used to schedule and regulate HVAC equipment.

If at least 80% of the building's total numbers of control system records are of a particular type, that type is chosen as the building's dominant flow control type. Note that the quantity of control system types must be used to scale the number of records. If the criteria are not met, the building's dominant flow control type is “Other Or Combination”. When a building does not report a Flow Control System, the default value is “Unknown”.

Heating Type

Heating types are equipment used to provide warmth to a building interior.

If the “Primary” flag is True for a particular heating type, that heating type is chosen as the building's dominant heating type. If those criteria are not met, or if the data is missing, the dominant heating type is decided according to heating capacity. If at least 80% of the building's heating capacity is of a particular heating type, that type is chosen. If the criteria is not met using the “Capacity” field, or if the data is missing, the criteria is then tested using the “Percent of Total Installed Capacity” field. If the criteria are still not met, or if the data is missing, the algorithm tests according to “Percent of Floor Area Served”. If at least 80% of the floor area served is of a particular heating type, that type is chosen. Note that the “Heating Type Quantity” field must be used to scale the capacities and floor areas. If the criteria are still not met, or if the data is missing, the next step is to check the number of heating records reported for the building. If there is only one heating type, that heating type is assigned as the dominant type. Finally, if none of the above criteria are met, the building's dominant heating type is “Other Or Combination”. When a building does not report a Heating Type, the default value is “Unknown”.

Cooling Type

Cooling types are equipment used to remove heat from a building interior.

If at least 80% of the building's cooling capacity is of a particular cooling type, that cooling type is chosen as the building's dominant cooling type. If the criteria is not met using the “Capacity” field, or if the data is missing, the criteria is then tested using the “Percent of Total Installed Capacity” field. If the criteria are still not met, or if the data is missing, the dominant facility is according to “Percent of Floor Area Served”. If at least 80% of the floor area served is of a particular cooling type, that cooling type is chosen as the dominant type. Note that the “Quantity” field must be used to scale the capacities and floor areas. If the criteria are still not met, or if the data is missing, the next step is to check the number of cooling types reported for the building. If there is only one, that cooling type is the dominant cooling type. Finally, if none of the above criteria are met, the building's dominant cooling type is “Other Or Combination”. When a building does not report a Cooling Type, the default value is “Unknown”.

Heating Fuel

Heating fuels are the fuels used by the heating equipment.

If the “Primary” flag is true for a particular heating fuel, that fuel is chosen as the building's dominant heating fuel. If that criterion is not met, or if the data is missing, the dominant heating type is decided according to capacity. If at least 80% of the building's heating capacity is of a particular heating fuel, that fuel is selected as the dominant heating fuel. If the criteria is not met using the “Heating Fuel Capacity”, or if the data is missing, the criteria is then tested using the “Percent of Total Installed Capacity” field. If the criteria are still not met, or if the data is missing, the dominant type is based on “Percent of Floor Area Served”. If at least 80% of the floor area served is of a particular heating fuel, that fuel becomes the dominant heating fuel. Note that the “Quantity” field must be used to scale the capacities and floor areas. If the criteria are still not met, or if the data is missing, the next step is to check the number of heating fuel records reported for the building. If there is only one, that heating fuel is chosen. Finally, if none of the above criteria are met, the building's heating fuel is “Other Or Combination”. When a building does not report a Heating Fuel, the default value is “Unknown”.

Wall Type

The wall type is the material composition of building exterior.

If at least 80% of the building's wall area is of a particular wall type, that type is chosen as the building’s dominant wall type. If the criteria is not met using the “Wall Area” field, or if the data is missing, the criteria is then tested using the “Percentage of Total Wall Area” field. If the criteria are still not met, or if the data is missing, the next step is to check the number of wall records reported for the building. If there is only one, that wall type is the dominant wall type.

Finally, if none of the above criteria are met, the building's wall type is “Other Or Combination”. When a building does not report a Wall Type, the default value is “Unknown”.

Roof/Ceiling Type

The roof/ceiling type is the roofing material or roof structure.

If at least 80% of the building's roof area is of a particular roof type, that type is chosen as the building's roof type. If the criteria is not met using the “Roof Area” field, or if the data is missing, the criteria is then tested using the “Percentage of Total Roof Area” field. If the criteria are still not met, or if the data is missing, the algorithm checks the number of roof records reported for the building. If there is only one, that roof type is chosen. Finally, if none of the above criteria are met, the building's roof type is “Other Or Combination”. When a building does not report a Roof/Ceiling Type, the default value is “Unknown”.

Window Glass Type

The window glass type is the window glass used at the building.

If at least 80% of the building's window area is of a particular window glass type, that type is chosen as the building's window glass type. If the criteria is not met using the “Window Area” field, or if the data is missing, the algorithm checks the number of window records reported for the building. If there is only one, that window glass type is chosen. Finally, if none of the above criteria are met, the building's window glass type is “Other Or Combination”.

Window Glass Layers

The window glass layers is the number of window layers installed the building.

If at least 80% of the building's window area is of a particular window glass layer, that layer is chosen as the building's window glass layer. If the criteria is not met using the “Window Area” field, or if the data is missing, the algorithm checks the number of window records reported for the building. If there is only one, that window glass layer is chosen. Finally, if none of the above criteria are met, the building's window glass layer is “Other Or Combination”.

Energy Use Calculations

The energy total is the sum of all individual time series readings during the most recent 365 days for which all streams have data. Separate totals are calculated for electricity, fuel (i.e., all non-electric fuel streams), site, and source energy. Energy use intensity (EUI) is calculated by dividing the energy total by the building's floor area.

Measurement Period

Ideally, a building should have 365 consecutive days of measured data, but buildings with slightly less data are accepted. Measurement periods of at least 358 days, an allowance of a single week in a year, but no more than 365 days are accepted. The measurement period may have gaps of one week or less spread out within the period, as long as all the gaps combined are two weeks or less. Otherwise, the energy total should be empty. The BPD team believes that allowing one weeks gaps (either within the year or at the beginning/end of the year) is an appropriate requirement, considering the impact of one versus two weeks of hot or cold weather on monthly energy use.

The measurement periods for electric, fuel, and site and source energy should be treated separately. The measurement period for electric energy is the most recent period in which all electric streams have energy data. The measurement period for fuel energy is the most recent period in which all fuel streams have energy data. The measurement period for site and source energy is the most recent period in which all fuel streams have data. For example, if a building has electricity data for the years 2011 and 2013 and fuel data for the years 2011 and 2012, then the total electricity calculated is for the year 2013, and the total fuel is for the year 2012, and the site and source calculated is for the year 2011.

Scaling Energy Totals

If the measurement period is less than 365 days, or if the measurement period has gaps, the energy total should be scaled up. For example, if a building has 360 days of data with 3 one day gaps, the energy total should be scaled up by a factor of 365/357. This method was determined to be adequate because the gaps were small relative to the length of a year.

It is possible that the start and end of the energy measurement period does not coincide with the start and end time stamps of the individual readings. In such a case, the energy use during the individual reading period should be scaled down before it is added to the energy total according to the proportion of the individual reading period that overlaps with the energy measurement period. For example, if the measurement period is from January 1, 2013 through December 31, 2013, and one of the individual reading periods is from December 16, 2013 through January 15, 2014, then the energy use during that individual reading period should be scaled down by a factor of 16/31 before being added to the energy total.

Conversion Factors

The following multipliers should be used to convert energy data from various units to a common unit (kBtu).

Table 2. Conversion Factors to kBtu (EPA 2013)

Unit	**Multiplier**
Wh (Watt hours)	0.00341214
MWh (million Watt-hours)	3412.14
kBtu (thousand Btu)	1
therms	100
kWh (thousand Watt-hours)	3.41214
Btu (British thermal units)	0.001
MBtu (million Btu)	1000

The following multipliers should be used to convert energy data from various units to a common unit (kBtu), depending on the fuel.

Table 3. Conversion Factors to kBtu by Meter Type for the US and Canada (EPA 2013)

Unit	Fuel	Multiplier
Cubic Meters	Natural Gas	0.029166304
cf (cubic feet)	Natural Gas	1.03
cf (cubic feet)	Propane	2.57083
ccf (hundred cubic feet)	Natural Gas	102.9
kcf (thousand cubic feet)	Natural Gas	1029
MCF (million cubic feet)	Natural Gas	1029000
Gallons	Propane	91.65
Gallons	Bottled gas (LPG or propane)	91.65
Gallons	Liquid Propane	91.65
Gallons	Fuel Oil	138.69
Gallons	Fuel Oil No.1	138.69
Gallons	Fuel Oil No.2	138.69
Gallons	Fuel Oil No.4	138.69
Gallons	Fuel Oil No.5 and No.6	149.69
Gallons	Diesel	138.69
Gallons	Kerosene	135
Lbs. (pounds)	District Steam	1.19
Lbs. (pounds)	Wood	7.69
Lbs. (pounds)	Coal	12.465
Lbs. (pounds)	Coal (anthracite)	12.545
Lbs. (pounds)	Coal (bituminous)	12.465
KLbs. (thousand pounds)	District Steam	1190
KLbs. (thousand pounds)	Wood	7690

KLbs. (thousand pounds)	Coal	12465
KLbs. (thousand pounds)	Coal (anthracite)	12545
KLbs. (thousand pounds)	Coal (bituminous)	12465
MLbs. (million pounds)	District Steam	1190000
MLbs. (million pounds)	Wood	7690000
MLbs. (million pounds)	Coal	12465000
MLbs. (million pounds)	Coal (anthracite)	12545000
MLbs. (million pounds)	Coal (bituminous)	12465000
Tons	District Steam	2380
Tons	Wood	15380
Tons	Coal	24930
Tons	Coal (anthracite)	25090
Tons	Coal (bituminous)	24930
Ton Hours	District chilled water	12

The following multipliers should be used to convert from site energy to source energy, depending on the fuel.

Table 4. U.S. Source-Site Ratio Calculations for Electricity (EPA 2013)

Fuel	**Multiplier**
Electricity	3.14
Electricity (Renewable)	1
Natural Gas	1.05
Natural Gas (Renewable)	1.05
Fuel Oil	1.01
Fuel Oil No.1	1.01
Fuel Oil No.2	1.01
Fuel Oil No.4	1.01

Fuel Oil No.5 and No.6	1.01
District Steam	1.2
District hot water	1.2
District chilled water	1
Solar hot water	1
Bottled gas (LPG or propane)	1.01
Propane	1.01
Liquid Propane	1.01
Kerosene	1.01
Diesel	1.01
Coal	1
Coal (anthracite)	1
Coal (bituminous)	1
Coke	1
Wood	1
Other Or Combination	1

Net Energy Measurements

The energy data in the BPD is reported as net energy uses (i.e., net use = consumption - generation). For nearly all buildings in the BPD, generated energy is known to be either zero or very small, so net energy use is approximately equal to consumed energy. For a small portion of buildings, it is not known whether the energy data provided was net energy or consumed energy, so the BPD team thought it safest to assume net energy.

Negative and Zero Energy

Site and source energy totals cannot be negative or zero. Individual fuel readings and fuel energy totals cannot be negative. Individual fuel readings or fuel energy totals can be zero. Individual electric readings can be zero or negative, but electric energy totals cannot be negative.

Miscellaneous

If a building has no electricity data, the energy totals for site and source are set as empty, but the energy total for fuel may not be empty. Likewise, if a building has no fuel data, the energy totals for site and source and set as empty, but the energy total for electricity may not be empty.

If a building has neither electric nor fuel data, then all energy totals should be empty.

Energy streams with “Add to Total Energy Use” set to false should not be added to any energy totals, nor should energy streams with Reading Type equal to anything other than “Total”, “Other”, or “Unknown”.

Some buildings have energy streams with Fuel = “Other Or Combination”, meaning it is not known whether this data is electric only, fuel only, or some combination of electric and fuel. If all of a building's energy streams have Fuel = “Other Or Combination”, then the energy should be added to the energy totals for site, but the energy totals for electric, fuel, and source should be empty. If a building has at least one energy stream without Fuel = “Other Or Combination”, then streams with Fuel = “Other Or Combination” should be added to the energy totals for fuel, site, and source (with a site to source conversion factor of one).

In summary the cleansing step of the process includes, checking for minimum data requirements, checking for allowed values, reviewing for consistency, and adding rules to account for dataset specific issues.

4. Handling Duplicated Data

Building records can appear multiple times in different data sources especially if the building is actively monitoring their energy consumption. A building can appear in municipal benchmarking records or in energy efficiency programs such as ENERGY STAR Portfolio Manager. Checking for duplicates in buildings ensures that a single building is not represented twice in the database. The BPD team checks for and removes duplicated buildings from the database both within each dataset and across multiple datasets.

Within a dataset, the primary source of duplicated records is when a contributor provides newly updated data (e.g., 2011 data was updated with 2012 data). If the building IDs remain the same from year to year, the old data is replaced with the new data (unless the old data has more fields). If the IDs are not consistent from year to year, the records are assumed to be different and are both kept (unless they are determined to be duplicates when checking across multiple datasets).

When checking for duplicates across multiple datasets, the BPD team takes the safe stance of assuming records are duplicates unless they can be proven to be unique. A pair of records is considered a duplicate if they have the same values for all the fields listed in **Table 5**.

Table 5. Duplicate matching criteria

Field	**Tolerance**
Facility Type	exact
Postal Code	exact
City	exact
State	exact

Climate zone	exact
Gross Floor Area	within 5%
Site Year	within 3 months
Site EUI	within 5%
Source Year	within 3 months
Source EUI	within 5%
Electric Year	within 3 months
Electric EUI	within 5%
Fuel Year	within 3 months
Fuel EUI	within 5%
Year Built	exact
Number of People	within 10%
Operating Hours	within 10 hours per week
Energy Star Rating (if year matches)	exact

If either record in the pair is missing data for the field being compared, the pair is considered to be a possible match; the pair is only considered a non-match when both records have data for a field and the data are different. When a pair is considered a match, the records with the lower score is marked for removal. The score is computed by taking the weighted sum of the number of fields without missing data. The scores are chosen to reflect a preference for certain fields over others: Facility Type, Postal code, and Gross Floor Area have a weight of 5, State and Climate have a weight of 3, Site EUI and Year have a weight of 2.5, Electric and Source EUI and Year have a weight of 1.5, and all other fields have a weight of 1.

In some cases, three or more buildings can all match each other. In these cases, the one building with the highest score is kept and the two or more buildings with the lowest scores are removed.

Lessons Learned

Based on the experience of performing data preparation for multiple datasets, several important tips, tricks, and lessons have helped the BPD data team to ensure the data quality of the database.

General Data Preparation

- Using a flexible schema allows the inclusion of sparse or exhaustive datasets. The schema should also adapt to trends in building technologies, for example, the increased adoption of on-site power generation in buildings.
- Request that data providers send documentation for their data terms and definitions. Use this document when translating input into BEDES. If there is any doubt about the asset data accuracy, ask for clarification from the data providers or leave the data field blank.
- While BEDES seeks to capture as much data as possible, sometimes value judgments need to be made and the data may lose some degree of specificity. For example, Portfolio Manager uses building type “K-12 School” which is mapped to “Education - Uncategorized” since there is no specific equivalent in BEDES.
- Inconsistencies in building identification codes, floor area, etc. among different raw files, make data preparation more time-consuming. A dataset commonly can provide building characteristic and energy data from different surveys. The building identification used in the surveys may be different from each other; the BPD team then looks at the floor area and location information to match buildings with asset information with their corresponding energy data.
- Data is dirty until proven clean. The BPD throws out roughly 20% of buildings provided by data contributors due to identified data quality issues such as failing the minimum data checklist or high estimated energy use values. There is a trade-off between data quality and quantity: while including more buildings in your analysis may reduce uncertainty, problematic data will lessen confidence in your results.

Energy Data Verification

- Multifamily buildings, can have inconsistent whole-building energy use. For instance, the heating fuel consumption is reported for the whole building, but electricity consumption is reported only for common areas. The current process is to sum the electricity for common areas and the tenant spaces to achieve whole building energy use.
- Data with inconsistent time intervals are difficult to aggregate. A complete discussion of the process can be found in the Energy Use Calculations section.
- Buildings with sub-meters or meters out of service can complicate calculation of energy totals. As long as sub-meter and master meter data are properly flagged or there is adequate information on building energy consumption, the buildings are included in the database.

Future Improvements

As the BPD incorporates more datasets with varying data formats and quality, the data preparation methods described here will need to evolve. Some of these future developments include: properly representing on-site energy generation and mixed-use buildings in the database. For data mapped using BEDES fields, the team have scripts that can translate from BEDES to the schema format.

As the BPD schema changes, new validation rules are needed to accommodate additional fields or a modification of the fields. These changes could be renaming equipment or building types or adjusting the in-range checks to reflect changing building technology efficiencies. The rules should also be flexible in order to cater to new features added to the BPD tool. The LBNL team will update this document accordingly, and flag any changes or additions.

Although automation has many advantages, the data preparation process still requires a human in the loop, most especially during mapping. Finding persistent errors in a dataset, default values in the contributor's schema, are usually verified with a phone call or email exchange. Matching the contributed data with BEDES is an activity that cannot be automated because of the variation in data formats and schema. A data processor still needs to write a data mapping script specific to the dataset.

Conclusion

The data preparation process is an important step between raw data submitted by data providers and the data populating the Buildings Performance Database. This process takes considerable time and effort, but is essential for providing a dataset that can be analyzed properly. The process requires conversion of raw data from data providers to BEDES format, deriving data from multiple data fields, and validating entries as summarized in Figure 3. The data preparation process will evolve as more building information is made available.

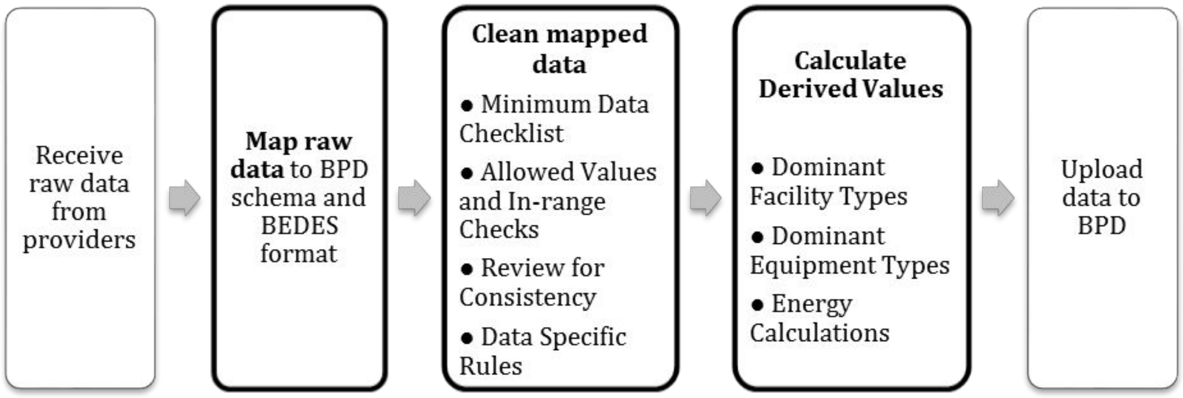

Receive raw data from providers
Map raw data to BPD schema and BEDES format
Clean mapped data
● Minimum Data Checklist
● Allowed Values and In-range Checks
● Review for Consistency
● Data Specific Rules
Calculate Derived Values
● Dominant Facility Types
● Dominant Equipment Types
● Energy Calculations
Upload data to BPD

Figure 3. Detailed summary of the BPD Data Preparation Process

References

Department of Energy. “Building Energy Data Exchange Specification (BEDES)”, Last modified April 18, 2015. http://www.energy.gov/eere/buildings/building-energy-data-exchange-specification-bedes

California Energy Commision. “California Commercial End-Use Survey (CEUS)”, Last modified March 23, 2007. http://www.energy.ca.gov/ceus/

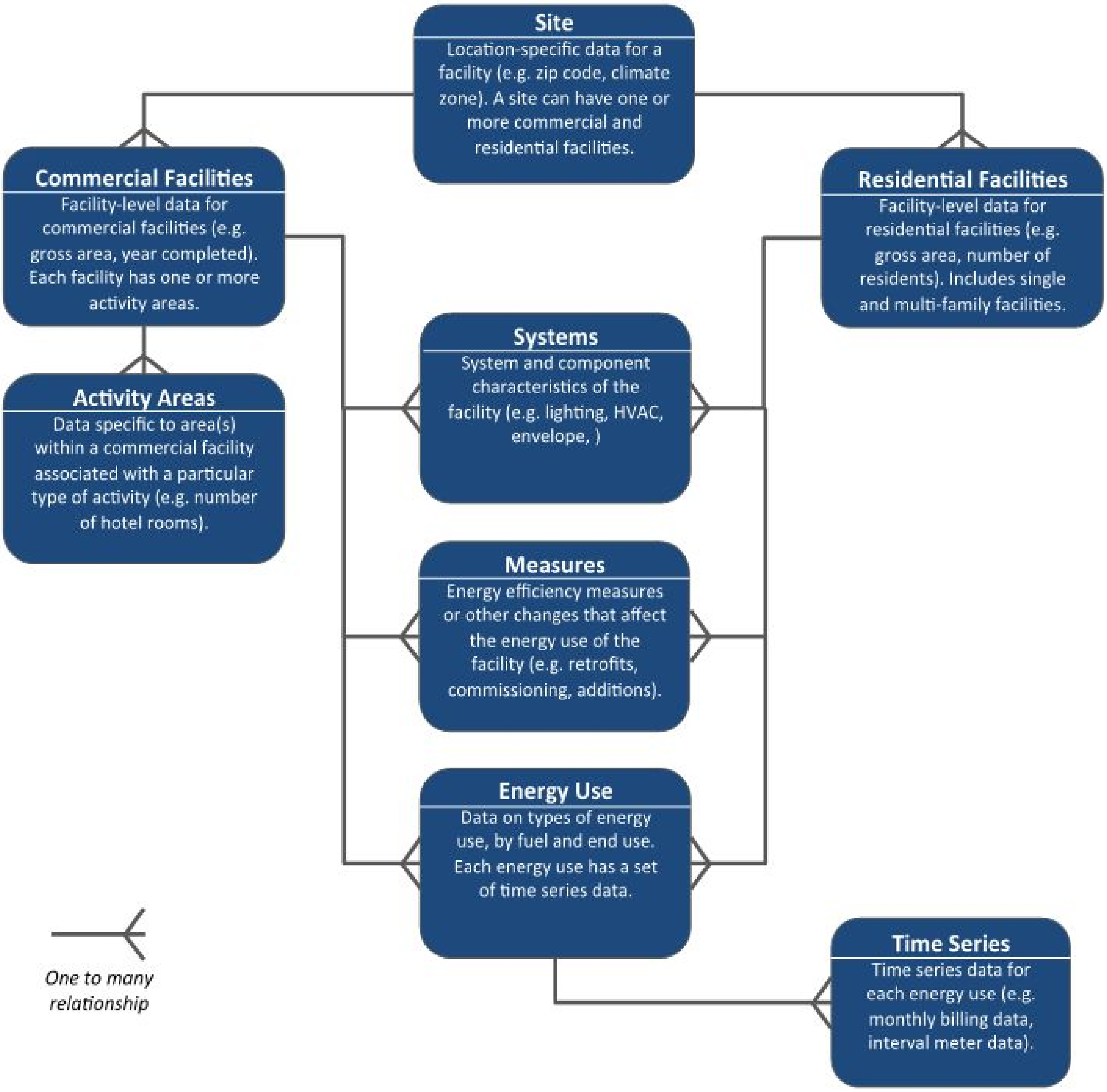
Energy Information Administration. “Energy Information Administration (EIA) - Commercial Buildings Energy Consumption Survey”, Last modified March 30, 2012. http://www.eia.gov/consumption/commercial/

Energy Information Administration. “Residential Energy Consumption Survey (RECS) - Energy Information Administration”, Last modified March 7, 2013 http://www.eia.gov/consumption/residential/

ENERGY STAR. “Portfolio Manager Technical Reference: Thermal Conversion Factors”, Last modified July 2013. http://www.energystar.gov/buildings/tools-and-resources/portfolio-manager-technical-reference-thermal-conversion-factors

Appendix A: Buildings Performance Database Data Schema

The basic structure of the data in the BPD is shown below:


Site
Location-specific data for a facility (e.g. zip code, climate zone). A site can have one or more commercial and residential facilities.
Commercial Facilities
Facility-level data for commercial facilities (e.g. gross area, year completed). Each facility has one or more activity areas.
Residential Facilities
Facility-level data for residential facilities (e.g. gross area, number of residents). Includes single and multi-family facilities.
Activity Areas
Data specific to area(s) within a commercial facility associated with a particular type of activity (e.g. number of hotel rooms).
Systems
System and component characteristics of the facility (e.g. lighting, HVAC, envelope,)
Measures
Energy efficiency measures or other changes that affect the energy use of the facility (e.g. retrofits, commissioning, additions).
Energy Use
Data on types of energy use, by fuel and end use. Each energy use has a set of time series data.
Time Series
Time series data for each energy use (e.g. monthly billing data, interval meter data).
One to many relationship

Appendix B: The BPD Cleansing and Validation Rules

The data types are similar to data types in programming language. ALPHANUMERIC values contain letters upper or lowercase A-Z and/or numerals 0-9. CHAR contains only letters, while DOUBLE values are decimal numerals and INTEGER() values are signed whole numbers. A CONSTRAINED LIST is list of valid values found in BEDES and BOOLEAN are values that are either TRUE or FALSE.

Site			
Field	**Data Type**	**Allowed Values**	**In-range checks** **Consistency checks**
Source Facility ID	ALPHANUMERIC		Must be unique
City	CHAR		City corresponds to Postal Code
State	CONSTRAINED LIST	List	State corresponds to the Postal Code
Postal Code	INTEGER(5)	List	00210 - 99950
County	CHAR		
Country	CHAR	List	
Climate zone	CONSTRAINED LIST	List	Climate zone corresponds to the Postal Code
Elevation	DOUBLE	-282 – 20320 feet	Negative elevation only allowed in CA, and LA; Outside of Alaska, the highest elevation is 14433 feet
Site Type	CONSTRAINED LIST	List	
Number of Facilities	INTEGER	>=1	
Complex Type	CONSTRAINED LIST	List	Field applies only if Number of Facilities >1
School District	CHAR		
eGRID Region	CONSTRAINED LIST	List	
Tax Floor Area	DOUBLE	100 - 7 million square feet	Equal to within 3% of the sum of the facilities' gross floor areas.

Residential Facility			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Residential Facility Type	CONSTRAINED LIST	List	
Complete Total Fuel	BOOLEAN	T or F	
Year Completed	INTEGER(4)	1600 - present	
Year Occupied	INTEGER(4)	1600 - present	>= Year Completed
Operator Type	CONSTRAINED LIST	List	
Surroundings	CONSTRAINED LIST	List	
Orientation	CONSTRAINED LIST	List	
Building Footprint Area	DOUBLE	100 - 1 million square feet	<= Gross Floor Area If Number of Floors = 1, then Building Footprint Area = Gross Floor Area.
Footprint Shape	CONSTRAINED LIST	List	
Perimeter	DOUBLE	40 - 30,000 feet	
Gross Floor Area	DOUBLE	100 - 1 million square feet	
Net Floor Area	DOUBLE	100 - 1 million square feet	<= Gross Floor Area
Rentable Floor Area	DOUBLE	0 - 1 million square feet	<= Gross Floor Area
Occupied Floor Area	DOUBLE	0 - 1 million square feet	<= Gross Floor Area
Percentage of Common Space	DOUBLE	0 – 100%	<= ((Gross Floor Area-Rentable Floor Area)/Gross Floor Area)*100
Lighted Floor Area	DOUBLE	0 - 1 million square feet	<= Gross Floor Area
Heated Floor Area	DOUBLE	0 - 1 million square feet	<= Gross Floor Area
Cooled Floor Area	DOUBLE	0 - 1 million square feet	<= Gross Floor Area
Unconditioned Floor Area	DOUBLE	0 - 1 million square feet	<= Gross Floor Area <= Gross Floor Area - maximum(Cooled Floor Area, Heated Floor Area)
Basement Floor Area	DOUBLE	0 - 600,000 square feet	<= Gross Floor Area*0.6 Else if Floors Below Ground <= 1, then Basement Floor Area <=

			Building Footprint Area. Else, Basement Floor Area >= Building Footprint Area.
Basement Heated Floor Area	DOUBLE	0 - 600,000 square feet	<= Gross Floor Area*0.6 <= Basement Floor Area
Basement Cooled Floor Area	DOUBLE	0 - 600,000 square feet	<= Gross Floor Area*0.6 <= Basement Floor Area
Attic Floor Area	DOUBLE	0 - 600,000 square feet	<= Gross Floor Area*0.6 <= Building Footprint Area
Attic Heated Floor Area	DOUBLE	0 - 600,000 square feet	<= Gross Floor Area*0.6 <= Attic Floor Area
Attic Cooled Floor Area	DOUBLE	0 - 600,000 square feet	<= Gross Floor Area*0.6 <= Attic Floor Area
Garage Floor Area	DOUBLE	0 - 400,000 square feet	<= Gross Floor Area*0.4
Garage Heated Floor Area	DOUBLE	0 - 400,000square feet	<= Gross Floor Area*0.4 <= Garage Floor Area
Garage Cooled Floor Area	DOUBLE	0 - 400,000 square feet	<= Gross Floor Area*0.4 <= Garage Floor Area
Volume	DOUBLE	600 - 130 million cubic feet	>= Gross Floor Area*6
Conditioned Building Volume	DOUBLE	600 - 130 million cubic feet	<= Volume >= Gross Floor Area*6
Number of Dwelling Units	INTEGER	1 – 900	
Aspect Ratio	DOUBLE	1 – 20	
Number of Floors	INTEGER	1 – 90	If Building Footprint Area < Gross Floor Area, then Number of Floors > 1 = Floors Above Ground + Floors Below Ground
Floors Above Ground	INTEGER	0 – 90	<= Number of Floors
Floors Below Ground	INTEGER	0 – 10	<= Number of Floors
Number of Residents	INTEGER	1 – 2500	(Gross Floor Area/Number of Residents) >= 78 < Number of Bedrooms*3
Number of Bedrooms	INTEGER	> 0	<= Number of rooms

Number of Complete Baths	INTEGER	> 0	<= Number of rooms
Number of Half Baths	INTEGER	> 0	<= Number of rooms
Number of Rooms	INTEGER	1 - 2500	
Building Certification Type	CONSTRAINED LIST - multiple instances	List	
Building Certification Value	ALPHANUMERIC - multiple instances		
Certification Year	INTEGER(4) - multiple instances	1990 - present	
Year of Last Remodel	INTEGER(4)	1600 - present	> Year Completed
Facility Number	ALPHANUMERIC		<= Number of Facilities in Site
Year Completed Range	CHAR		Ranges from 1600 - present
Number of Floor Plans	INTEGER	1 – 900	<= Number of Dwelling Units
Floor Plan Type	CONSTRAINED LIST	List	
Number of Units per Floor Plan	INTEGER - multiple instances	1 – 900	<= Number of Dwelling Units
Facility Tax Floor Area	DOUBLE	100 - 1 million square feet	= Gross Floor Area +/- Gross Floor Area*0.03
Enclosed Floor Area	DOUBLE	100 - 1 million square feet	>= Gross Floor Area
Conditioned Floor Area	DOUBLE	0 - 1 million square feet	<= Gross Floor Area >= Heated Only Floor Area + Cooled Only Floor Area
Semi-conditioned Floor Area	DOUBLE	0 - 1 million square feet	<= Conditioned Floor Area
Heated Only Floor Area	DOUBLE	0 - 1 million square feet	<= Conditioned Floor Area
Cooled Only Floor Area	DOUBLE	0 - 1 million square feet	<= Conditioned Floor Area
Heated and Cooled Floor Area	DOUBLE	0 - 1 million square feet	<= Conditioned Floor Area
Non-Enclosed Floor Area (w/roof)	DOUBLE	0 - 1 million square feet	

Open Floor Area (w/o roof)	DOUBLE	0 - 1 million square feet	
Facility Height	DOUBLE	6 – 900 feet	
Floor-to-Floor Height	DOUBLE	6 – 300 feet	<= Facility Height
Number of Floors in Housing Unit	INTEGER	1 – 10	
Number of Conditioned Floors	INTEGER	1 – 90	<= Number of Floors or Number of Floors in Housing Unit
Number of Conditioned Floors Above Ground	INTEGER	0 – 90	<= Number of Conditioned Floors
Numbers of Conditioned Floors Below Ground	INTEGER	0 – 10	<= Number of Conditioned Floors
Household Type	CONSTRAINED LIST	List	
Ownership Status	CONSTRAINED LIST	List	
Government Subsidized Housing	BOOLEAN	T or F	
Occupant Income Range	CONSTRAINED LIST	List	
Occupant Education	CONSTRAINED LIST	List	
Resident Population Type	CONSTRAINED LIST	List	
Number of Adults	INTEGER	1 – 2500	<= Number of Residents
Number of Children	INTEGER	0 – 2500	<= Number of Residents
Number of Non-bedrooms	INTEGER	> 0	< Number of rooms

Commercial Facility			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Complete Total Energy	BOOLEAN	T or F	
Primary Facility Type	CONSTRAINED LIST	List	
Year Completed	INTEGER(4)	1600 - present	
Year Occupied	INTEGER(4)	1600 - present	>= Year Completed
Owner Type	CONSTRAINED LIST	List	

Percent Occupied by Owner	DOUBLE	0 - 100	
Operator Type	CONSTRAINED LIST	List	If Percent Occupied by Owner = 100, then Operator Type = “Owner”
Surroundings	CONSTRAINED LIST	List	
Orientation	CONSTRAINED LIST	List	
Building Footprint Area	DOUBLE	100 - 2 million square feet	<= Gross Floor Area If Number of Floors = 1, then Building Footprint Area = Gross Floor Area.
Footprint Shape	CONSTRAINED LIST	List	
Perimeter	DOUBLE	40 - 30,000 feet	
Gross Floor Area	DOUBLE	100 - 7 million square feet	
Net Floor Area	DOUBLE	100 - 7 million square feet	<= Gross Floor Area
Rentable Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area
Occupied Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area
Lighted Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area
Heated Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area >= Sum of Heated Floor Area in all Activity Areas
Cooled Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area >= Sum of Cooled Floor Area in all Activity Areas
Unconditioned Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area <= Gross Floor Area - maximum(Cooled Floor Area, Heated Floor Area)
Volume	DOUBLE	600 - 263 million cubic feet	>= Gross Floor Area*6
Conditioned Building Volume	DOUBLE	0 - 263 million cubic feet	<= Volume >= Gross Floor Area*6
Aspect Ratio	DOUBLE	1 – 20	
Number of Floors	INTEGER	1 – 110	If Building Footprint Area < Gross Floor Area, then Number of Floors > 1 = Floors Above Ground + Floors Below Ground >= Maximum Number of Floors in all Activity Areas

Floors Above Ground	INTEGER	0 – 110	<= Number of Floors >= Maximum Floors Above Ground in all Activity Areas
Floors Below Ground	INTEGER	0 – 10	<= Number of Floors >= Maximum Floors Below Ground in all Activity Areas
Number of Occupants	INTEGER	0 - 35,000	>= Sum of Number of Occupants in all Activity Areas
Building Certification Type	CONSTRAINED LIST - multiple instances	List	
Building Certification Value	ALPHANUMERIC - multiple instances		
Certification Year	INTEGER(4) - multiple instances	1990 – present	
Year of Last Remodel	INTEGER(4)	1600 – present	> Year Completed
Number of Activity Areas	INTEGER	1 – 100	
Facility Number	ALPHANUMERIC		<= Number of Facilities in Site
NAICS Code	CHAR		
Ownership Status	CONSTRAINED LIST	List	
Facility Tax Floor Area	DOUBLE	100 - 7 million square feet	= Gross Floor Area +/- Gross Floor Area*1.03
Enclosed Floor Area	DOUBLE	100 - 7 million square feet	>= Gross Floor Area >= Sum of Enclosed Floor Area in all Activity Areas
Conditioned Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area >= Heated Only Floor Area + Cooled Only Floor Area >= Sum of Conditioned Floor Area in all Activity Areas
Semi-conditioned Floor Area	DOUBLE	0 - 7 million square feet	<= Conditioned Floor Area
Heated Only Floor Area	DOUBLE	0 - 7 million square feet	<= Conditioned Floor Area
Cooled Only Floor Area	DOUBLE	0 - 7 million square feet	<= Conditioned Floor Area
Heated and Cooled Floor Area	DOUBLE	0 - 7 million square feet	<= Conditioned Floor Area
Non-Enclosed Floor Area (w/roof)	DOUBLE	0 - 3 million square feet	>= Sum of Non-Enclosed Floor Area (w/roof) in all Activity Areas

Open Floor Area (w/o roof)	DOUBLE	0 - 3 million square feet	>= Sum of Non-Enclosed Floor Area (w/o roof) in all Activity Areas
Basement Floor Area	DOUBLE	0 - 5 million square feet	<= Gross Floor Area*0.05 Else if Floors Below Ground <= 1, then Basement Floor Area <= Building Footprint Area. Else, Basement Floor Area >= Building Footprint Area.
Facility Height	DOUBLE	6 - 1,450 feet	
Floor-to-Floor Height	DOUBLE	6 – 50 feet	<= Facility Height
Number of Conditioned Floors	INTEGER	0 – 110	<= Number of Floors
Number of Conditioned Floors Above Ground	INTEGER	0 – 110	<= Number of Conditioned Floors <= Number of Floors
Numbers of Conditioned Floors Below Ground	INTEGER	0 – 10	<= Number of Conditioned Floors <= Number of Floors

Activity Area			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Activity Type	CONSTRAINED LIST	List	
Gross Floor Area	DOUBLE	0 - 7 million square feet	
Heated Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area
Cooled Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area
Number of Floors	INTEGER	1 – 110	= Floors Above Ground + Floors Below Ground
Floors Above Ground	INTEGER	0 – 110	<= Number of Floors
Floors Below Ground	INTEGER	0 – 10	<= Number of Floors
Floor Height	DOUBLE	6 – 300 feet	
Number of Units	INTEGER	1 - 3,000	
Seating Capacity	INTEGER	0 - 1,000	
Number of Occupants	INTEGER	0 - 35,000	

Number of Guest Rooms	INTEGER	1 – 6500	
Number of Licensed Beds	INTEGER	0 - 2,500	
Average Weekly Operating Hours	DOUBLE	0 – 168 hours	
Average Weeks per Year in Use	DOUBLE	0 – 52	
Activity Area Name	ALPHANUMERIC		
Metered Space	BOOLEAN	T or F	
Exterior Entrance to the Public	BOOLEAN	T or F	
Enclosed Floor Area	DOUBLE	100 - 8 million square feet	>= Gross Floor Area
Non-Enclosed Floor Area (w/roof)	DOUBLE	0 - 3 million square feet	
Open Floor Area (w/o roof)	DOUBLE	0 - 3 million square feet	
Conditioned Floor Area	DOUBLE	0 - 7 million square feet	<= Gross Floor Area <= Heated Floor Area + Cooled Floor Area >= Heated Only Floor Area + Cooled Only Floor Area
Office Air-Conditioned	BOOLEAN	T or F	
Office Heated	BOOLEAN	T or F	
Senior Care Facility Total Resident Capacity	INTEGER	1 - 1,500	
Occupant Density	DOUBLE		
Hours per Day Guests on Site	DOUBLE	0 – 24 hours	
Number of Months in Use	DOUBLE	0 – 12	
Average Occupancy Percentage	DOUBLE	0 – 100%	
Room Density	DOUBLE		
Number of Businesses	INTEGER	0 – 800	
Weekly Hours of Parking Access	DOUBLE	0 – 168 hours	
K - 12 School - Open Weekends	BOOLEAN	T or F	

Floor area of full-service spas	DOUBLE	0 - 134,000 square feet	
Floor area of gym/fitness center	DOUBLE	0 - 200,000 square feet	
Dorm has Computer Lab	BOOLEAN	T or F	
Hospital - Tertiary Care?	BOOLEAN	T or F	

Measures			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Measure List	CONSTRAINED LIST - Multi-select	List	
Scope	DOUBLE	0 – 100	
Description	CHAR		
Start Date	DATE	After 01/01/1600	Format MM/DD/YYYY.
End Date	DATE	After 01/01/1600	End Date should be within 10 years of Start Date. Format MM/DD/YYYY.
Life	INTEGER	0 – 100 years	
First Cost	DOUBLE	0 - 10 million US dollars	
O&M Cost	DOUBLE	0 - 10 million US dollars	
Funding from Rebates	DOUBLE	0 - 10 million US dollars	<= First Cost
Funding from Tax Credits	DOUBLE	0 - 10 million US dollars	<= First Cost
Lot Configuration	CONSTRAINED LIST	List	
Simple Payback	DOUBLE	0 – 100 years	
Cost Effectiveness Screening Method	CONSTRAINED LIST	List	
O&M Cost Annual Savings	DOUBLE	0 - 10 million US dollars/year	
Total Job Hours	DOUBLE	0 – 8760 hours	
Annual Savings Estimate (Cost)	DOUBLE	0 - 10 million US dollars/year	
Annual Savings Estimate (Energy)	DOUBLE	0 - 10 million Btu/year	
Annual Savings (Cost)	DOUBLE	0 - 10 million US dollars/year	

Annual Savings (Energy)	DOUBLE	0 - 10 million Btu/year	
Implementation Status	CONSTRAINED LIST	List	
Number of Permits Replaced	INTEGER	0 - 10,000	
Number of Staff Members Trained	INTEGER	0 - 35,000	
Work Performed By	CONSTRAINED LIST	List	

Lighting			
Field	**Data Type**	**Allowed Values**	**In-range checks**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Lighting Type	CONSTRAINED LIST	List	
Ballast Type	CONSTRAINED LIST	List	
Lighting Control Type	CONSTRAINED LIST	List	
Installed Power	DOUBLE	1 - 175 million kW	< 25*Gross Floor Area of Facility
Percentage of Total Installed Power	DOUBLE	0 – 100%	Sum of all Percentage of Total Installed Power <= 100
Percentage of Total Floor Area Served	DOUBLE	0 – 100%	If Outside Lighting = F, then Percentage of Total Floor Area <= (Lighted Floor Area/Gross Floor Area)*100
Outside Lighting	BOOLEAN	T or F	
Lighting Efficacy	DOUBLE	1 – 683 lm/W	
Specular Reflectors	BOOLEAN	T or F	
Percent lit when open	DOUBLE	0 – 100%	
Percent lit when closed	DOUBLE	0 – 100%	
Lighting Daily Hours	CONSTRAINED LIST	0 – 24 hours	

Air Distribution			
Field	**Data Type**	**Allowed Values**	**In-range checks**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table

Configuration	CONSTRAINED LIST	List	
Flow Control	CONSTRAINED LIST	List	
Duct Configuration	CONSTRAINED LIST	List	
Heating Source	CONSTRAINED LIST	List	
Cooling Source	CONSTRAINED LIST	List	
Preheat Source	CONSTRAINED LIST	List	
Humidification	CONSTRAINED LIST	List	
Dehumidification	CONSTRAINED LIST	List	
Quantity	INTEGER	1 - 10,000	
Size	DOUBLE	0 - 345,000 cfm	
Power	DOUBLE	1 - 175 million kW	< 25*Gross Floor Area of Facility
Year of Manufacture	INTEGER(4)	1885 - present	>= 10 - facility Year Completed
Primary	BOOLEAN	T or F	if Percent of Total Installed Capacity >= 80, then "T"
Percent of Total Installed Capacity	DOUBLE	0 – 100%	Sum of all Percentage of Total Installed Capacity <= 100
Percent of Floor Area Served	DOUBLE	0 – 100%	
Static Pressure Reset Control	BOOLEAN	T or F	
Supply Air Temperature Reset Control	BOOLEAN	T or F	
Efficiency	DOUBLE	1 – 200 W/cfm 0 – 100%	
Efficiency Unit	CONSTRAINED LIST	List	If there is an Efficiency value, then there must be an Efficiency Unit
Fan Motor Efficiency	DOUBLE	1 - 100%	
Economizer	CONSTRAINED LIST	List	
Minimum Outside Air Percentage	DOUBLE	0 – 100%	
Fan Control Type	CONSTRAINED LIST	List	

Heat Recovery Type	CONSTRAINED LIST	List	
Heat Recovery Efficiency	DOUBLE	1 – 100%	
Duct Insulation	CONSTRAINED LIST	List	
Duct Sealing	CONSTRAINED LIST	List	
Duct Location	CONSTRAINED LIST	List	
Duct Insulation R-Value	INTEGER(2)	0 – 15 ft^{2}-ºF-h/Btu-in	
Static Pressure	DOUBLE	0 - 3,000 Pa	
Zone Count	INTEGER	1 - 1,000	
Fan Placement	CONSTRAINED LIST	List	
Flow Configuration	CONSTRAINED LIST	List	
Duct Type	CONSTRAINED LIST	List	
Bucket Type	CONSTRAINED LIST	List	
Duct Pressure Test Leakage (cfm)	DOUBLE	0 - 345,000 cfm	
Duct Pressure Test Leakage (Percentage)]	DOUBLE	0 - 100 %	

Heating			
Field	**Data Type**	**Allowed Values**	**In-range checks**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Heating Type	CONSTRAINED LIST	List	
Burner Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Quantity	INTEGER	1 - 1,000	
Capacity	DOUBLE	0 - 26 million Btu/hour	
Capacity Unit	CONSTRAINED LIST	List	
Year of Manufacture	INTEGER(4)	1885 – present	
Primary	BOOLEAN	T or F	If Percent Floor Area >= 80%, then T
Efficiency	DOUBLE	AFUE: 0- 100, COP: 0 - 20, SEER: 0 -30, HSPF: 0 - 20, Percent: 0-100	

Efficiency Unit	CONSTRAINED LIST	List	If there is an Efficiency value, then there must be an Efficiency Unit
Percent of Total Installed Capacity	DOUBLE	0 – 100%	Sum of all Percent of Total Installed Capacity in building <= 100
Percent of Floor Area Served	DOUBLE	0 – 100%	
Hot Water Reset Control	CONSTRAINED LIST	List	
Control Type	CONSTRAINED LIST	List	
Distribution Type	CONSTRAINED LIST - Multi-select	List	
Vacuum Return System	BOOLEAN	T or F	
Input Capacity	DOUBLE		Dependent on equipment type
Output Capacity	DOUBLE	0 – 800 horsepower	<= Input Capacity
Location	CONSTRAINED LIST	List	
Draft Type	CONSTRAINED LIST	List	
Zone Count	INTEGER	1 - 1,000	<= Sum of all Zones in Air Distribution
Exclusive to this Facility	BOOLEAN	T or F	

Zonal Heating			
Field	**Data Type**	**Allowed Values**	**In-range checks**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Type	CONSTRAINED LIST	List	
Control Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Year of Manufacture	INTEGER(4)	1600 - present	
Quantity	INTEGER	1 - 1,000	
Percent of Total Installed Capacity	DOUBLE	0 – 100%	Sum of all Percent of Total Installed Capacity in building <= 100
Percent of Floor Area Served	DOUBLE	0 – 100%	

Input Capacity	DOUBLE		Dependent on equipment type
Output Capacity	DOUBLE		<= Input Capacity
Efficiency	DOUBLE	0-100	
Efficiency Unit	CONSTRAINED LIST	List	If there is an Efficiency value, then there must be an Efficiency Unit

Cooling			
Field	**Data Type**	**Allowed Values**	**In-range checks**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Cooling Type	CONSTRAINED LIST	List	
Chiller Compressor Type	CONSTRAINED LIST	List	
Chiller Cooling Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Quantity	INTEGER	1 - 1,000	
Capacity	DOUBLE	1-13,000	
Capacity Unit	CONSTRAINED LIST	List	
Efficiency	DOUBLE	kW/ton: < 1.5 AFUE: 0- 100, COP: 0 - 20, SEER: 0 -30, HSPF: 0 - 20, EER: 0 – 40 Percent: 0-100	
Efficiency Unit	CONSTRAINED LIST	List	If there is an Efficiency value, then there must be an Efficiency Unit
Year of Manufacture	INTEGER (4)	1885 - present	
Primary	BOOLEAN	T or F	If Percent of Floor Area >= 80%, then T
Energy Star	BOOLEAN	T or F	If T, the Year Manufacture >= 1995
Percent of Total Installed Capacity	DOUBLE	0 – 100%	Sum of all Percent of Total Installed Capacity in building <= 100%

Percent of Floor Area Served	DOUBLE	0 – 100%	
Chilled Water Reset Control	CONSTRAINED LIST	List	
Cooling Tower Control Type	CONSTRAINED LIST	List	
Control Type	CONSTRAINED LIST	List	
Water-side Economizer	BOOLEAN	T or F	
Location	CONSTRAINED LIST	List	
Zone Count	INTEGER	1 - 1,000	<= Sum of all Zones in Air Distribution
Cooling Equipment Redundancy	BOOLEAN	T or F	

Zonal Cooling			
Field	**Data Type**	**Allowed Values**	**In-range checks**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Zonal Cooling Type	CONSTRAINED LIST	List	
Year of Manufacture	INTEGER(4)	1885 - present	
Control Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Quantity	INTEGER	1 - 1,000	
Percent of Total Installed Capacity	DOUBLE	0 – 100%	Sum of all Percent of Total Installed Capacity in building <= 100%
Percent of Floor Area Served	DOUBLE	0 – 100%	
Size Range	CONSTRAINED LIST	List	
Downstream Distribution	BOOLEAN	T or F	
Average Operating Hours	DOUBLE	0 – 24 hours	
Number of Months in Operation	DOUBLE	1 – 12 months	

Other HVAC

Field	Data Type	Allowed Values	In-range checks
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Other HVAC Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Quantity	INTEGER	1 - 1,000	
Capacity	DOUBLE		
Capacity Unit	CONSTRAINED LIST	List	If there is a Capacity value, then there must be an Capacity Unit
Year of Manufacture	INTEGER(4)	1885 - present	
Percent of Total Installed Capacity	DOUBLE	0 – 100%	Sum of all Percent of Total Installed Capacity in building <= 100
Percent of Floor Area Served	DOUBLE	0 – 100%	
Control Type	CONSTRAINED LIST	List	
Location	CONSTRAINED LIST	List	
Zone Count	INTEGER	1 - 1,000	<= Sum of all Zones in Air Distribution

Service Hot Water			
Field	**Data Type**	**Allowed Values**	**In-range checks**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Service Hot Water Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Quantity	INTEGER	1 - 1,000	
Size	DOUBLE	1-250,000 gallons	
Capacity	DOUBLE		Depends on equipment type
Capacity Unit			If there is an Capacity value, then there must be an Capacity Unit
Year of Manufacture	INTEGER(4)	1885 – present	

Energy Star	BOOLEAN	T or F	If T, then Year of Manufacture >= 2009
Control Type	CONSTRAINED LIST	List	
Efficiency	DOUBLE	AFUE: 0- 100, COP: 0 - 20, SEER: 0 -30, HSPF: 0 - 20, EER: 0 – 40 Percent: 0-100	
Efficiency Unit	CONSTRAINED LIST	List	If there is an Efficiency value, then there must be an Efficiency Unit
Storage Tank Insulation R-Value	INTEGER(2)	0 – 30 ft^2-°F-h/Btu-in	
Storage Tank Insulation Thickness	DOUBLE	0 – 10 inches	
Percent of Total Installed Capacity	DOUBLE	0 – 100%	Sum of Percent of Total Installed Capacity in building <= 100
Location	CONSTRAINED LIST	List	
Setpoint Temp	DOUBLE	100 – 200 °F	
Residential Temperature Setting	CONSTRAINED LIST	List	

General Controls & Operations			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Regular HVAC Maintenance	CONSTRAINED LIST	List	
Programmable Thermostats Used	BOOLEAN	T or F	
Standard Thermostats	BOOLEAN	T or F	
Heat Lowered	CONSTRAINED LIST	List	
AC Adjusted	CONSTRAINED LIST	List	
Occupied Day Setting	DOUBLE	0; 55 – 100 °F	
Unoccupied Day Setting	DOUBLE	0; 55 – 100 °F	
Sleeping Hours Setting	DOUBLE	0; 55 – 100 °F	

Task Lighting In Use	BOOLEAN	T or F	
Percent Controlled by EMCS	DOUBLE	0 – 100%	
Primary HVAC Control Strategy	CONSTRAINED LIST	List	
HVAC Systems Controlled by EMCS	CONSTRAINED LIST - Multi-select	List	
Individual HVAC Control	CONSTRAINED LIST	List	
Percent of rooms controlled by thermostatic radiator valves	DOUBLE	0 – 100%	
Percent of rooms controlled by electronic zone valves with thermostats	DOUBLE	0 – 100%	
Percent of rooms controlled by temperature sensors	DOUBLE	0 – 100%	
Percent of rooms controlled by demand control ventilation	DOUBLE	0 – 100%	
Lighting reduced during off hours	BOOLEAN	T or F	
Percent Area Covered by Occupancy Sensors	DOUBLE	0 – 100%	
Percent Area Covered by Vacancy Sensors	DOUBLE	0 – 100%	
Percent Area Covered by Photo Sensors	DOUBLE	0 – 100%	
Percent Area Covered by Timers	DOUBLE	0 – 100%	
Percent Area Covered by Advanced Controls	DOUBLE	0 – 100%	
AC Replaced in Last 10 Years	BOOLEAN	T or F	
Number of Coils Replaced	INTEGER	0 - 1,000	
Number of Air Handlers Replaced	INTEGER	0 - 1,000	
Number of Guest Meals	INTEGER	0 - 35,000	

Quantity of Laundry	INTEGER	0 - 35 million	

Wall			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Exterior Wall Type	CONSTRAINED LIST	List	
Exterior Wall Color	CONSTRAINED LIST	List	
Wall R-Value	INTEGER(2)	0 – 60 ft^2-°F-h/Btu-in	
Percentage of Total Wall Area	DOUBLE	0 – 100%	Sum of all Percentage of Wall Area should not be more than 100
Wall Insulation Type	CONSTRAINED LIST	List	
Wall Insulation Thickness	DOUBLE	0 – 10 in	
Basement wall insulation thickness	DOUBLE	0 – 10 in	
Tightness	CONSTRAINED LIST	List	
Wall Area	DOUBLE	0 - 175 million square feet	

Roof-Ceiling			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Roof Type	CONSTRAINED LIST	List	
Roof Color	CONSTRAINED LIST	List	
Deck Type	CONSTRAINED LIST	List	
Roof R-Value	INTEGER(2)	0 – 60 ft^2-°F-h/Btu-in	
Percentage of Total Roof Area	DOUBLE	0 – 100%	
Roof Insulation Type	CONSTRAINED LIST	List	
Roof Insulation Thickness	DOUBLE	0 – 10 in	
Attic/Ceiling Type	CONSTRAINED LIST	List	
Attic/Ceiling R-value	INTEGER(2)	0 – 60 ft^2-°F-h/Btu-in	

Attic/Ceiling Insulation thickness	DOUBLE	0 – 10 in	
Attic/Ceiling Insulation Type	CONSTRAINED LIST	List	
Radiant Barrier	BOOLEAN	T or F	
Roof Slope	CONSTRAINED LIST	List	
Roof Area	DOUBLE	100 - 2 million square feet	Sum of all Roof Areas must be >= Building Footprint Area of the facility
Percent of Roof Terraces	DOUBLE	0 – 100%	
Terrace R-Value	INTEGER(2)	0 – 60 ft^{2}-°F-h/Btu-in	
Attic Access Location	CONSTRAINED LIST	List	

Fenestration			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Window Glass Type	CONSTRAINED LIST	List	
Operable Windows	BOOLEAN	T or F	
Windows Gas Filled	BOOLEAN	T or F	
Window Glass Layers	CONSTRAINED LIST	List	
Window R-value	INTEGER(2)	0 – 12 ft^{2}-°F-h/Btu-in	
Solar Heat Gain Coefficient (SHGC)	DOUBLE	0 - 1	
Window Visible Transmittance	DOUBLE	0 - 1	
Window to Wall Ratio	DOUBLE	0 -1	
Window Frame Type	CONSTRAINED LIST	List	
Exterior Shading Type	CONSTRAINED LIST	List	
Exterior Shading Orientation	CONSTRAINED LIST	List	
Interior Shading Type	CONSTRAINED LIST	List	
Skylights	BOOLEAN	T or F	
Windows Weather-Stripped	BOOLEAN	T or F	

Exterior Door Type	CONSTRAINED LIST	List	
Doors Weather-Stripped	BOOLEAN	T or F	
Window Area	DOUBLE	1 - 250,000 square feet	
Ground Floor Only	BOOLEAN	T or F	
Percent Vision Glazing	DOUBLE	0 – 100%	
Year of Last Window Replacement	INTEGER(4)	1600 – present	> Year Completed of facility
Percent of Window Area Shaded	DOUBLE	0 – 100%	
Percent Skylight Area	DOUBLE	0 – 100%	
Skylight SHGC	DOUBLE	0 – 1	
Skylight Visible Transmittance	DOUBLE	0 – 1	
Number of Exterior Doors	INTEGER	1 - 1,000	

Foundation			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Ground Coupling	CONSTRAINED LIST	List	
Perimeter Insulated	BOOLEAN	T or F	
Floor Insulation Thickness	DOUBLE	0 – 10 in	
Floor R-Value	INTEGER(2)	0 – 60 ft^2-°F-h/Btu-in	
Floor Construction Type	CONSTRAINED LIST	List	
Carpet	BOOLEAN	T or F	
Plumbing Penetration Sealing	BOOLEAN	T or F	

IT System			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
IT System Type	CONSTRAINED LIST	List	
Quantity	INTEGER	1 - 28,000	

Electrical Plug Intensity	DOUBLE	1 - 175,000,000 Watts/sq.ft	< 25*Gross Floor Area of Facility
Energy Star	BOOLEAN	T or F	If T, then Year Manufacture >= 1992
UPS System Redundancy	BOOLEAN	T or F	
Density	DOUBLE	units/1,000 square feet	

Process Load			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Process Load Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Quantity	INTEGER	1 - 28,000	
Installed Power	DOUBLE	1 - 175,000,000 Watts	< 25*Gross Floor Area of Facility
Efficiency Value	DOUBLE	0 – 100	
Efficiency Unit	CONSTRAINED LIST	List	
Capacity	DOUBLE	1 - 175,000,000 Watts	
Capacity Unit	CONSTRAINE LIST	List	If there is Capacity value, then there must be a Capacity Unit

Cooking			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Cooking Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Capacity	DOUBLE	1 - 175,000,000 Watts	
Capacity Unit	CONSTRAINED LIST	List	If there is Capacity value, then there must be a Capacity Unit
Quantity	INTEGER	0 - 1,000	

Refrigeration			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Refrigeration Type	CONSTRAINED LIST	List	
Quantity	INTEGER	0 - 1,000	
Size	DOUBLE	1 - 175,000,000 cubic feet	
Year of Manufacture	INTEGER(4)	1950 - present	
Energy Star	BOOLEAN	T or F	If T, then Year Manufacture >= 1996
Doors	CONSTRAINED LIST	List	
Density	DOUBLE		

Dishwasher			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Dishwasher Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Year of Manufacture	INTEGER(4)	1950 - present	
Energy Star	BOOLEAN	T or F	If T, then Year of Manufacture >= 1996
Quantity	INTEGER	1 - 2,500	
Loads Per Week	DOUBLE	1 - 17,500	

Laundry			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Laundry Type	CONSTRAINED LIST	List	
Fuel	CONSTRAINED LIST	List	
Year of Manufacture	INTEGER(4)	1950 - present	
Energy Star	BOOLEAN	T or F	If T, then Year of Manufacture >= 1997

Quantity	INTEGER	1 - 2,500	

Conveyance			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Conveyance System Type	CONSTRAINED LIST	List	
Quantity	INTEGER	1 – 100	
Control Type	CONSTRAINED LIST	List	

On-Site Generation			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
On-Site Generation Type	CONSTRAINED LIST	List	
Fuel Generated	CONSTRAINED LIST	List	
Quantity	INTEGER	1 – 100	
Capacity	DOUBLE		Depends on equipment
Capacity Unit	CONSTRAINED LIST	List	If there is Capacity value, then there must be a Capacity Unit
Annual Operation Hours	DOUBLE	0 – 8760 hours	

Energy Storage			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Energy Storage Type	CONSTRAINED LIST	List	
Quantity	INTEGER	1 - 100	
Capacity	DOUBLE		Depends on equipment
Capacity Unit	CONSTRAINED LIST	List	If there is Capacity value, then there must be a Capacity Unit

Pool			
Field	**Data Type**	**Allowed Values**	**In-range check**
Source Facility ID	ALPHANUMERIC		Match Source Facility ID in Site table
Pool Type	CONSTRAINED LIST	List	
Quantity	INTEGER	1 – 40	
Heated	BOOLEAN	T or F	
Fuel	CONSTRAINED LIST	List	
Pool Volume	DOUBLE	100 - 6.5 million	
Number of Pool Pumps	INTEGER	0 – 100	
Control Type	CONSTRAINED LIST	List	
Indoor	BOOLEAN	T or F	
Pool Area	DOUBLE	20 - 150,000 square feet	

Energy Use			
Field	**Data Type**	**Allowed Values**	**In-range checks**
Source Facility ID	ALPHANUMERIC		
Energy Use ID	Integer	>=1	
Fuel	CONSTRAINED LIST	List	
Add to Total	BOOLEAN	T or F	IT Energy (UPS) should be False unless otherwise stated.
End Use Type	CONSTRAINED LIST	List	
Units	CONSTRAINED LIST	List	
Interval Type	CONSTRAINED LIST	List	
Reading Time Zone Code	CHAR(3)	List	
Reading Type	CONSTRAINED LIST	List	
Summer Peak	DOUBLE		<= Maximum Time Series Reading
Winter Peak	DOUBLE		<= Maximum Time Series Reading
Rate Structure	CHAR		
Metering Configuration	CONSTRAINED LIST	List	
Emissions Factor	DOUBLE	0 – 200 kg CO2e/MMBtu	

Fuel Interruptibility	CONSTRAINED LIST	List	
Shared Energy System	CONSTRAINED LIST	List	

Time Series			
Field	**Data Type**	**Allowed Values**	**In-range checks**
Source Facility ID	ALPHANUMERIC		
Energy Use ID	Integer	>=1	
Start Time Stamp	DATE	1/1/1600 0:00 – Present	Cannot be earlier than year built. Format MM/DD/YYYY hh:mm
End Time Stamp	DATE	Start time + Length of Interval Type	If interval type is annual, metering time must be 365 days +/- 7 days. If interval type is monthly, must have at least 11 months of data.
Reading	DOUBLE	>=0	EUI = [0, 2000] kBtu/ft^2
Interval Peak	DOUBLE	>=0	<= Reading
Interval Minimum	DOUBLE	>=0	<= Reading <= Interval Peak

Appendix C: Minimum Data Requirements

Field Name	Description
SITE DATA	
Source Facility ID	Unique alphanumeric identifier for each facility.
COMMERCIAL FACILITY DATA	
Source Facility ID	Should match Source Facility ID's in Site Data.
Primary Facility Type	Specific activity designated for majority of the property. Accepted activity types are specified in BEDES.
Gross Floor Area	Floor area in square feet.
ACTIVITY AREA DATA **(For Commercial Facilities)**	
Source Facility ID	Should match Source Facility ID's in Site Data.
Activity Type	Specific activity designated for majority of the property. Accepted activity types are specified in BEDES.
Gross Floor Area	Floor area, in square feet, associated with the activity area.
RESIDENTIAL DATA	
Source Facility ID	Should match Source Facility ID's in Site Data.
Residential Facility Type	Specific residential type, refer to BEDES for accepted Residential Types.
Gross Floor Area	Floor area, in square feet, associated with the residential type.
ENERGY USE DATA	
Source Facility ID	Should match Source Facility ID's in Site Data.
Fuel	Type of fuel used in the building record.
Units	Unit of measurement for each fuel stream.

TIME SERIES DATA	
Source Facility ID	Should match Source Facility ID's in Energy Use Data.
Start Time Stamp	Timestamp that marks the beginning of the fuel record.
End Time Stamp	Timestamp that marks the end of the fuel record.
Reading	Numerical value of the reading.

Additional Energy Record Requirements

a) Each building needs to have a continuous year of energy records for major fuel types.

b) Energy data for different fuel types should be reported within the same time period, this is important when considering seasonal fuel usage especially during the winter;

c) Other than facilities with energy generated on-site with net metering, energy readings must be greater than or equal to 0; and

d) Calculated annual site EUI for a single building or facility should be greater than or equal to 1 kBtu/ft^2/year and less than or equal to 1,000 kBtu/ft^2/year. Calculated annual source EUI for a single building or facility should be greater than or equal to 1 kBtu/ft^2/year and less than or equal to 3,140 kBtu/ft^2/year. Calculated annual electric and fuel EUI for a single building or facility should be greater than or equal to 0 and less than or equal to 1,000 kBtu/ft^2/year. The limits were determined based on ENERGY STAR's Portfolio Manager where the maximum median source EUI is 600 kBtu/ft^2/year, in Retail buildings but the maximum reported EUI is a little more than 1,000 kBtu/ft^2/year reported for Office buildings.

Appendix D: Pseudo-code for Error Checking

The error-checking rules are turned into Python scripts to increase efficiency and expedite the data preparation process. The following pseudo-code shows the process flow for the error-checking script:

```
-   Read in data parsed into respective tables.
-   Find buildings that are missing from “Site”, “Time Series” and “Energy Use”
    tables.
-   Find buildings that are missing from “Commercial Facility” or “Residential
    Facility” tables or are in both Facility tables.
-   Find energy uses for each building that are missing from either “Time Series”
    or “Energy Use” tables.
-   Find buildings that are missing from “Commercial Facility” and “Activity Area”
    tables.
-   Read in the list of error checking rules.
-   For each table:
        o   Check for default values. Default values appear for more than 30% of the
            buildings in following fields: “Year Completed”, “Gross Floor Area”,
            “Number of Occupants”, and “Average Weekly Operating Hours”.
        o   Find the error checking rules relevant to the table.
                ▪   Check for mandatory fields.
                ▪   Check for percent values that are not in the correct format.
                ▪   Check that numerical values are in range of the defined threshold.
                ▪   Check that numerical values are consistent with other values, e.g.
                    “Enclosed Floor Area” is less than or equal to the “Gross Floor
                    Area”.
                ▪   Check that zipcode values are properly formatted.
                        ●   Check the correct “City”, “State”, and “Climate zone”
                            matches the given “Postal code”.
                ▪   Run through the error rules that do not fit in the general
                    categories, e.g. “Elevation” for buildings from different states.
                ▪   Check that enumerated values are in the BEDES list.
        o   Log the errors.
-   Fix errors logged from error checking.
        o   Delete buildings missing mandatory fields, missing from required tables,
            missing postal code, or with out of range gross floor area.
        o   Add missing fields to a table.
        o   Delete entries if the value is out of range, not in the BEDES lists, or
            other violations of error checking rules.
        o   Fix the format of percent values to be between 1-100.
-   Fill in empty numerical field columns with “No Value” and others with
    “Unknown”.
-   Print error checked and fixed tables to “Error Checked” directory.
```

Appendix E: Pseudo-code for Deriving Values

The derived values are turned into Python scripts to be consistent across datasets and reduce the error in calculating values. The following pseudo-code shows the process flow for the derived values script:

```
-   Read in the “Site” table.
-   Rename columns: "Source Facility ID" to "id", "Postal Code" to "zip_code", "Climate zone" to "climate", "State" to "state", "City" to "city".
-   Read in the available tables with asset information e.g., “Heating” and “Lighting”.
-   Calculate the characteristics-related derived values.
      o   Calculate the dominant “Facility Type”.
      o   Calculate the “Operating Hours”.
      o   Calculate building information: “Year Built”, “Floor Area”, and “Building Class”.
      o   Calculate “Certification Values”.
      o   Calculate dominant “Wall Type” and average “Wall R-value”.
      o   Calculate dominant “Roof Type”.
      o   Calculate dominant “Window Glass Type” and “Window Glass Layers”.
      o   Calculate dominant “Lighting Type”.
      o   Calculate dominant “Air Flow Control”.
      o   Calculate dominant “Cooling Type”.
      o   Calculate dominant “Heating Type”.
-   Fill in the empty values with “No Value” for numerical values and “Unknown” for other data types.
-   Read in the “Time Series”, “Energy Use”, “Commercial Facility”, and “Residential Facility”.
-   Calculate energy-related derived values for each building.
-   Calculate the site, source, electricity and fuel energy use intensity.
      o   Convert all electricity and fuel stream readings to kBtu.
      o   Find the time span based on the start and end time stamps.
      o   If the time span is within the 365-day threshold and meets the other criteria, return the total energy stream and the year of record.
      o   Check the flags and the fuel streams meet the requirements.
            ▪   Check “Complete Energy Flag”, “Complete Electricity Flag”, and “Complete Fuel Flag” are True.
            ▪   Check data stream for a given fuel is not empty.
      o   Calculate the respective EUIs. Divide the total energy consumption with the gross floor area.
      o   Check calculated energy values are in range.
      o   Log errors.
-   Print “Buildings” file and log file to “Final” directory.
```