

ELCAP Data Assembly and Conversion Project: Report on File Contents

December 14, 2012

Regional Technical Forum

This page left blank.

Principal Investigators:

Aaron Jenniges

Ryan Kilkenny

Niko Drake-McLaughlin

Matei Perussi

The Cadmus Group, Inc.

This page left blank.

Table of Contents

[Executive Summary 1](#_Toc342924878)

[Summary of Key Findings 1](#_Toc342924879)

[Proposed Database Platform and Schema 2](#_Toc342924880)

[ELCAP Data Files 4](#_Toc342924881)

[ELCAP Raw Hourly Load Data Files 4](#_Toc342924882)

[Site Level Hourly Meter Data 4](#_Toc342924883)

[Weather Station Hourly Data 7](#_Toc342924884)

[PNNL Data Files 9](#_Toc342924885)

[PNNL Site Files 9](#_Toc342924886)

[BRARES.SASDB.XPT 12](#_Toc342924887)

[CHAR.XFER.SASDB.XPT 12](#_Toc342924888)

[HRCP.DAILY.PADS.XPT and HRCP.WEEKLY.PADS.XPT 14](#_Toc342924889)

[NWPP.SASDB.XPT 14](#_Toc342924890)

[USRD.SASLIB.XPT 14](#_Toc342924891)

[PNNL AVAX Files 15](#_Toc342924892)

[Summarized Data 16](#_Toc342924893)

[Reports 16](#_Toc342924894)

[Database Structure and Schema 17](#_Toc342924895)

[Appendices 20](#_Toc342924896)

This page left blank.

# Executive Summary

In July 2012, the Regional Technical Forum (RTF) contracted with Cadmus to assemble the digital and written elements of the End Use Load and Consumer Assessment Program (ELCAP) data and convert them into a searchable database to be used by regional utilities and the RTF in evaluating hourly load shapes. Cadmus has reviewed all of the ELCAP files thoroughly. This report summarizes the contents of those files; it also documents which files are included in or excluded from the database. We provide information on the database platform and structure, and explain why the structure will make simple data analysis and extraction from the database efficient and convenient.

## Summary of Key Findings

The ELCAP data files are separated into four categories:

* Raw hourly site level meter and weather station data
* Pacific Northwest National Laboratory (PNNL) data, containing site characteristics
* Summary files containing aggregated tables based on summarized hourly load data
* Reports on analysis conducted on ELCAP load data.

There are 499 total files in the raw hourly site level meter and weather station data category, with each file corresponding to a unique site. Of these, 458 files contain hourly site level meter data, and 41 files contain hourly data from unique weather stations[[1]](#footnote-1). The hourly data in both the site level meter and weather station data files range from April 1, 1987 to August 31, 1989. The hourly site level meter files are grouped into six site type categories: residential, residential standards demonstration program, multifamily, commercial, commercial audit program, and purchase of energy savings program.

The PNNL data includes numerous characteristics files that can be linked to the site level meter files to provide detailed building information, such as location, site-to-weather-station mapping, climate zone, building type, and number of occupants. AVAX, a subset of files in the PNNL data category, contains 19 files stored in various formats. With the exception of one file containing a mapping of weather stations to the site level meter data, the AVAX files are saved in formats that would be difficult to import into a database structure. Most of the AVAX files appear to be duplicates of the characteristics files in the main PNNL category, and therefore are not included in the database.

The aggregated summary files are stored as either text files or Microsoft Excel workbooks. These files provide hourly site level meter data condensed into average load shapes by month, day type, and end use, and provide end-use intensities for various building types. They are included in the database, and are available to download through the Web portal.

Finally, Adobe Acrobat and Microsoft Word reports are included in the ELCAP raw data. These reports provide the results of analysis conducted on the ELCAP data and background information on the project. They are included in the database, and are available to download through the Web portal.

## Proposed Database Platform and Schema

Cadmus recommends that the database be stored in Microsoft SQL Server. Given that the ELCAP database tables contain millions of records, the size of the database is a critical factor in choosing a platform. Microsoft Access, for example, has a two-gigabyte limit on the total size of a database, making it insufficient to handle the ELCAP database. However, RDBMS (relational database management system) platforms such as Microsoft SQL Server do not have such limitations.

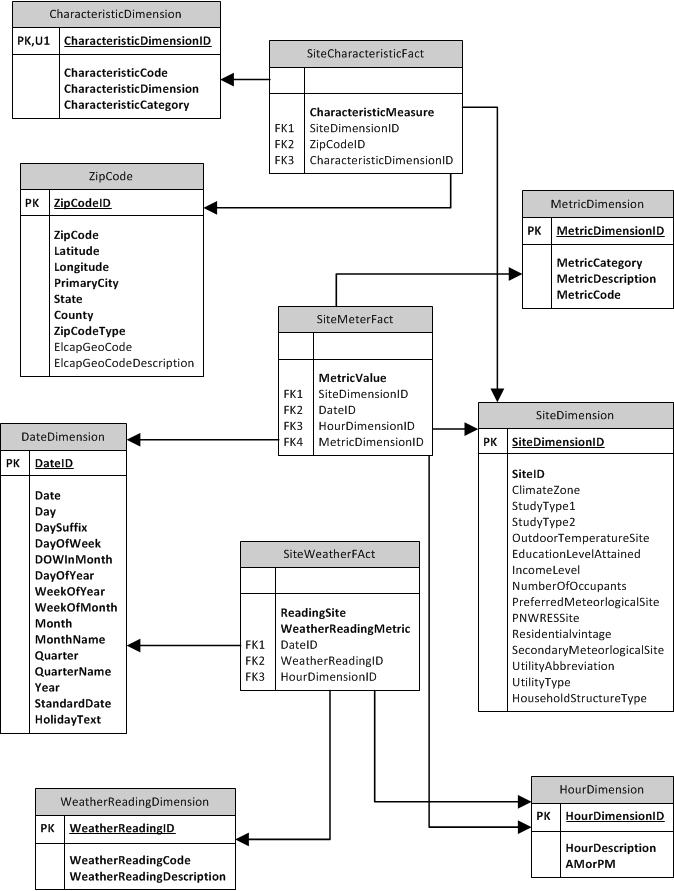
We have analyzed the hourly site level meter and weather station files, and the site level characteristics files, to determine the appropriate variables for linking the files. The primary variable used to link the hourly site level meter files to the site characteristics files is site number.

Through testing the linking of the files in SAS, we gathered information on the site level characteristics files, including which ones contain site number, which ones are duplicates of other files, what sectors they pertain to, and whether they will be imported into the database. Numerous records on many of the files contain site numbers that do not exist in the hourly site level meter data. Therefore, these records will be omitted from the database.

We implemented the structure and database schema diagram shown in Figure 1. The advantage of this multidimensional structure is it is compatible with cube analysis in SQL. Cubes provide an easy-to-use mechanism for querying data stored in a relational database, with quick and uniform response times. Users interact with the cube through an interface that allows them to easily manipulate the controls to obtain customized reports.

The primary tables in this schema are the fact tables: SiteMeterFact, SiteWeatherFact, and SiteCharacteristicFact. These tables contain the values (facts) for each possible combination of variables. Identification numbers in the other columns of the table provide the means for linking to the other tables in the schema, called dimensions. The dimensions provide additional information for each fact. An added advantage of this format is that dimension tables can include extraneous information that is useful for reporting on the facts, such as classifying dates into quarters or ZIP codes into counties.

Figure . Database Schema



# ELCAP Data Files

The ELCAP data files can be divided into four distinct categories. The pages that follow provide an in-depth description of the file contents for each of these categories:

* ELCAP raw hourly load data files
* PNNL data files
* Summary files
* Reports.

## ELCAP Raw Hourly Load Data Files

There are 499 total files in the raw hourly site level meter and weather station data category, with each file corresponding to a unique site. Of these, 458 files contain hourly site level meter data, and 41 files contain hourly weather data from unique weather stations.

### Site Level Hourly Meter Data

The hourly data in the 458 site level meter files range from April 1, 1987 to August 31, 1989. Eighteen commercial, twelve multifamily, four purchase of energy savings, and one residential site cannot be linked to site characteristics files because they are pseudo sites[[2]](#footnote-2), so there are 423 files included in the database. The sites not included in the database will be available to download in the web application. The hourly site level meter files are grouped into six site type categories, shown in Table 1 with the number of sites by category.

Table . Number of Sites by Category

|  |  |
| --- | --- |
| Category | Number of Sites |
| Commercial | 63 |
| Commercial Audit Program | 16 |
| Multifamily | 5 |
| Purchase of Energy Savings | 0 |
| Residential | 256 |
| Residential Standards Demonstration Program | 83 |

All of the site level meter files contain three common fields: site number, date, and hour. The site number is the unique identifier that can be used to link the hourly meter data for the site with other ELCAP files containing site characteristics. The hour represents the end of the interval (i.e., hour 1 indicates the hour from 12:00 a.m. to 1:00 a.m.). Most of the remaining fields in each file contain the hourly metered loads by end use, with each field pertaining to a unique end use. The number of fields varies from file to file, depending on how many end uses were metered for a particular site. A few extra fields in each file contain information on the quality of the data collected.

Appendix A provides a list of all the hourly site level meter file names, their sector (residential or commercial), category, number of observations, state date, end date, number of variable fields, percent of records that are considered good, bad, or missing, and whether the site is included in the database. Although all these files provide hourly data, the number of observations on each file varies because, for some of the sites, the metering data collection began after April 1, 1987, or ended before August 31, 1989. Some files contained records for the full date range, but were missing data for some days (percent shown in the Missing Data column). The observations with missing data were removed from the record count listed in the Number of Observations column of Appendix A.

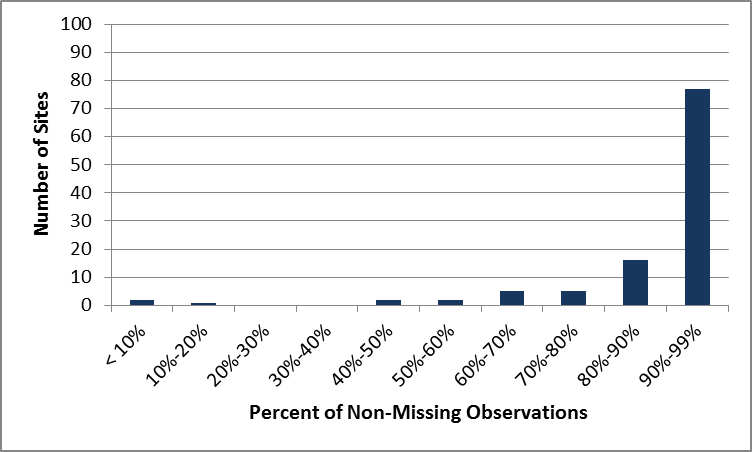
The maximum number of observations on a file is 22,216 (24 hours multiplied by the 884 days between April 1, 1987 and August 31, 1989). Figure 1 shows a frequency count of the percent of the maximum observations each site contains. More than 60% of the site files contain more than 90% of the possible number of hours. There are 11 site files that contain less than half of the possible number of hours.

Figure . Analysis of Amount of Hourly Data on Site Level Files

Figure 2 does not tell us the quality of the data, however. All of the site level meter files contain a field called DQ Level that provides an assessment of the quality of the metered data for each hour. Appendix A provides a site-by-site assessment of the quality of observations for each site in the Good Data and Bad Data columns. The percentages in these columns are calculated as percent of the maximum number of observations, including the missing records. An observation was considered good if it had a DQ Level between 1 and 5, and bad if it had a DQ Level between 6 and 9.

Figure 3 shows that the overall quality of the data is good. Almost 75% of the sites have no data that were rated as bad quality (and do not appear in the chart). Of the 25% of sites containing some bad data, 75% contained more than 90% of observations rated as good quality. Three files contained a low amount (less than 20%) of good quality data. All data, regardless of quality, is included in the database, with a corresponding data quality field attached to each observation to allow users access to the quality rating.

Figure 3. Analysis of Good Quality Data in Site Level Meter Files



Appendix C contains a comprehensive list of the data field abbreviations contained in the hourly site level meter file headers, along with a description of what type of data the fields contain. In the Number of Sites column, we have provided a count of the number of site files in which each field appears (out of the original 458 sites). We then assessed how frequently the fields contain data. The Percent Populated column shows the percentage the field is populated compared to all observations across all site level meter files. A low percentage indicates that data were collected for the particular field at few sites. Note that DQF and DQL contain data for all observations on all site level meter files, while A2D, LAB, and GFL rarely contain data. Regardless of this percentage, all fields are included in the database tables.

As described in more detail in the database design section of this report, we have added to the database the ability to group similar metered fields for reporting and analysis. Our designation of metered data fields into groups is represented in the Database Metric Category column of Appendix C. Table 2 provides a list of the categories and the fields assigned to each.

Table 2. Database Metric Categories

|  |  |
| --- | --- |
| Category | Fields |
| Data Quality | Analog To Digital Converter Reconstruction, Dq Flag, Dq Level, Logger Sumcheck |
| Food Preparation | Food Preparation |
| Heating/Cooling | Air Conditioning, Cooling, Duct Gas Flow, Heating, Heating (Proxy), Heating, Ventilation, And Air Conditioning, Mixed Hvac, Space Heating, Ventilation, Wooduse |
| Large Appliance | Clothes Drying, Clothes Washing, Dish Washing, Freezer, Refrigeration, Refrigerator |
| Lighting | Interior Lighting, Exterior Lighting, Lighting & Conveniences |
| Miscellaneous | Air To Heat Exchanger, Auxiliaries, Data Processing, General-Mixed, Laboratory, Material Handling, Other, Receptacles, Sanitation, Shop, Specialty #1, Specialty #2, Specialty #3, Unknown, Vertical Transport, Wood Stove Thermal Sensor |
| Totals | Building Total |
| Water Heater | Hot Water, Water Heating |
| Weather | Exterior Temp, Global Horiz Radiation, Interior Temp, Relative Humid, Wind Speed |

All of the hourly site level meter files are stored as SAS transport files, which are easily converted into SAS datasets and imported into a SQL database table. The data in these files are numeric. The metered values were recorded as average Watts over each hour. Cadmus has kept the data in average Watts in the database and has provided a conversion table so users can choose how to display the values in reports and exports (KWh, MWh, etc.).

### Weather Station Hourly Data

There are 41 unique weather station data files containing hourly weather data ranging from April 1, 1987 to August 31, 1989. Six of the weather station sites cannot be linked to the hourly site level meter data and have been excluded, so 35 weather stations are included in the database. The six excluded sites are given a tag of “No” in the Include in Database column of Appendix B.

As with the hourly site level meter files, all of the weather files contain three common fields: site number, date, and hour. The site number is a unique identifier that can be used to link the hourly weather station data with to the hourly site level meter files. The remaining fields in each file contain the collected hourly weather data. The number of fields varies from file to file, depending on how many fields for which data were collected.

Appendix B provides a list of all the weather station file names. Although all these files provide hourly data, the number of observations on each file varies slightly, because some files contained records that were missing data for some days. The observations with missing data were removed from the record count listed in the Number of Observations column of Appendix B. All of the weather files contained more than 90% of the maximum number of hours.

Some of the weather station sites provide actual readings of various weather metrics, while other sites provide TMY (typical meteorological year) averages. The Actual / TMY column in Appendix B shows which sites provide actual data and which provide TMY data. For the sites that provide actual data, the Station City / State column shows the major city closest to the weather station. We have included a mapping of the sites with actual data to the TMY sites for comparison, in the column labeled TMY Normal Station.

Appendix D contains a comprehensive list of the weather data field abbreviations contained in the weather data file headers, along with the translation of the abbreviations into ELCAP descriptions. In addition, we have provided a count of the number of files in which each field appears. Two fields, intermixed with the weather data and appearing on every file, are DQ Flag and DQ Level. These two fields provide information on the assessed quality of each observation in the weather station site files All data, regardless of quality, is included in the database, with a corresponding data quality field attached to each observation to allow users access to the quality rating.

All of the weather files are stored as SAS transport files, which are easily converted into SAS datasets and imported into a SQL database table. The data in these files are numeric.

## PNNL Data Files

The PNNL files include numerous characteristics files that can be linked to the hourly site level meter data by the site number, to provide detailed building information. There are 19 files categorized as AVAX files, which are a subset of the PNNL files.

### PNNL Site Files

Cadmus extracted 113 PNNL site datasets from six SAS transport files in the PNNL file directory. The other SAS transport files in the directory contain SAS formats. Table 3 provides an overview of the transport file names and their content.

Table 3. PNNL SAS Transport Files

|  |  |
| --- | --- |
| File Name | Number of Usable Data Sets |
| BRARES.CHARAC.FORMATS.XPT | formats |
| BRARES.FORMATS.XPT | formats |
| BRARES.SASDB.XPT | 9 |
| CHAR.XFER.SASDB.XPT | 87 |
| HRCP.DAILY.PADS.XPT | 1 |
| HRCP.WEEKLY.PADS.XPT | 1 |
| NWPP.SASDB.XPT | 4 |
| USRD.FMTLIB.XPT | formats |
| USRD.SASLIB.XPT | 16 |

Appendix E provides a list of all the datasets extracted from the SAS transport files, the dataset names, descriptions, and the source SAS transport file name. It also contains several columns providing additional details on the files. They are:

* Contains Site ID From Meter Data – “Yes” indicates that the file contains at least one site number for which site level meter data exist; “No” indicates the file does not contain site numbers for which site level meter data exist (and is not included in the database).
* Unique By Site ID – “Yes” indicates that each site number occurs on the file only one time; “No” indicates that at least one site number appears multiple times in the file making it difficult to link to the site level hourly meter data.
* Include in Database – “Yes” indicates that the file is included in the database; “No” indicates the file is not included in the database.
* Comments – Provides a description of why a file is not included in the final database or other relevant information about the file.
* Site Sectors – Provides a list of the sectors that pertain to the file, based on the site numbers contained in the file.
* Number of Meter Sites – Provides a count of unique site numbers from the meter data that appear on the file.
* Number of Sites – No Meter Data – Provides a count of the number of unique sites appearing on the file for which hourly meter data do not exist.
* Total Observations – Provides a count of the total number of observations in the file.
* Final Observations – Provides a count of the total number of observations the file has in the database (after we removed site numbers for which we do not have hourly meter data).

Table 4 provides a list of the PNNL files that will not be included in the database and the reason that they are excluded.

Table 4. PNNL Files Excluded from Database

| File Name | File Description | Contains Site ID from Meter Data | Unique By Site ID | Comment |
| --- | --- | --- | --- | --- |
| BILL83 | BILLING INFO - 1983 | Yes | Yes | Outdated Annual total kWh usage before meter installations |
| CHANNEL | CHANNEL | Yes | No | Meter sensor/connection location information. Many rows. |
| CHARAC | CHARACTERISTICS | Yes | Yes | Duplicate of Charac2 |
| CURC | (blank) | Yes | No | Duplicate of CRMATRIX |
| DAILY | (blank) | No | No | Aggregate data; no meter site numbers |
| HR10 | HRPNWRES | No | Yes | No meter site numbers |
| HR11 | HRMASTER | No | Yes | No meter site numbers |
| HR21 | HRPROJECTED | No | No | No meter site numbers |
| HR22 | HRCEILING\_PROJ | No | No | No meter site numbers |
| HR23 | HRDOORS\_PROJ | No | No | No meter site numbers |
| HR24 | HRDUCTS\_PROJ | No | No | No meter site numbers |
| HR25 | HRFLOORS\_PROJ | No | No | No meter site numbers |
| HR26 | HRGLASS\_PROJ | No | No | No meter site numbers |
| HR27 | HRWALLS\_PROJ | No | No | No meter site numbers |
| HR31 | HRACTUAL | No | Yes | No meter site numbers |
| HR32 | HRCEILING | No | No | No meter site numbers |
| HR33 | HRDOORS | No | No | No meter site numbers |
| HR34 | HRDUCTS | No | No | No meter site numbers |
| HR35 | HRFLOORS | No | No | No meter site numbers |
| HR36 | HRGLASS | No | No | No meter site numbers |
| HR37 | HRWALLS | No | No | No meter site numbers |
| HR41 | HRMEASURE | No | No | No meter site numbers |
| HR51 | HRCUSTOMER | No | No | No meter site numbers |
| HR52 | HROCCUPANTS | No | No | No meter site numbers |
| HR61 | HRBARRIER | No | No | No meter site numbers |
| HR71 | HRCOSTS\_PROJ | No | No | No meter site numbers |
| HR72 | HRCOSTS | No | No | No meter site numbers |
| HR81 | HRBIDS | No | No | No meter site numbers |
| NWPPBPA | HISTORIC BPA DATA, 1974-1987 | No | No Site ID | Historic Data |
| NWPPPRI | HISTORIC PUBLIC UTILIY DATA, 1974-1987 | No | No Site ID | Historic Data |
| NWPPPUB | HISTORIC PRIVATE UTILITY DATA, 1974-1987 | No | No Site ID | Historic Data |
| NWPPREG | HISTORIC REGULATORY DATA, 1974-1987 | No | No Site ID | Historic Data |
| PSEUDO | PSEUDO SITES MAP | Yes | No | Excluding pseudo sites |
| PXTOLOG | PXTOLOG | Yes | No | Duplicate of PX |
| RENAMED | (blank) | Yes | No | Duplicate of CRMATRIX |
| RIONE | (blank) | Yes | Yes | Duplicate of RI |
| RESMO | MONTHLY TOTAL LOAD AND WEATHER DATA | Yes | No | Same as RESMO2 where variables in common; contains less variables than RESMO2 |
| RESMO3 | MONTHLY TOTAL LOAD AND WEATHER DATA | Yes | No | Similar to RESMO2, but less complete |
| RESMO3B | MONTHLY TOTAL LOAD AND WEATHER DATA | Yes | No | Similar to RESMO2B, but less complete |
| RMATRIX | (blank) | Yes | Yes | Duplicate of Smatrix |
| UTILRATE | UTILITY\_RATES | No | No Site ID | Old rates |
| WEEKLY | (blank) | No | No | Aggregate Data |

Appendix F provides a list of the field names on each PNNL dataset, a description of the field, data type, and format. The format and format length columns indicate special SAS formats are applied to the data.

Numerous records on many of the PNNL characteristics files contain site numbers that do not exist in the hourly site level meter data. Therefore, these records will be omitted from the database. Appendix M provides a breakdown of the site numbers that appear on each of the PNNL characteristics files. The rows represent site number, with column A providing the sector for each site number (sheet is color-coded by sector). Each column represents a PNNL characteristics file. Any cell not color-coded but containing a site number represents a site number not represented in the hourly site level meter data.

## BRARES.SASDB.XPT

There were nine files on the BRARES.SASDB.XPT SAS transport file. Two datasets appear to have the same information: CHARAC and CHARAC2. The only difference between these two datasets is that CHARAC2 contains five additional fields. Given this, only the CHARAC2 file is included in the database. The CHARAC2 file is one of the primary building characteristics files for residential sites and can be mapped to the hourly site level meter files.

The remaining seven datasets on the BRARES.SASDB.XPT file can be split into two categories. Five of the datasets (with prefix RESMO) contain monthly total load and weather data, for some metered residential sites. The five files share many fields in common and appear to have some duplicate data. Appendix G contains a comprehensive list of fields in these files, and a matrix showing which fields are on which files. Our analysis revealed that RESMO is the same file as RESMO2, except it contains fewer variables than RESMO2. Therefore, RESMO will not be included in the database. Also, RESMO3 and RESMO3B appear to be the similar to RESMO2 and RESMO2B, except RESMO3 and RESMO3B contain more null values. Since RESMO2 and RESMO2B are more complete, they will be included in the database.

The other two datasets (SYPK and TWD) contain the results of data quality testing for residential sites. The SYPK data includes hourly data for up to three days per site (8/26/85, 1/7/86, or 8/8/86) for all end uses, with a data quality rating assigned for each hour. The TWD dataset shows the same type of data, except it includes one day per month for each site. The SYPK file appears to assign a data quality value to each hour of 0 to 4. The hourly data quality ranking fields on the TWD file appear to always be 0 or null. These files are not unique by site ID, so it is difficult to link them to the hourly site level meter files.

## CHAR.XFER.SASDB.XPT

The 87 datasets in the CHAR.XFER.SASDB.XPT file contain a variety of site characteristics, including details on building tenants in the commercial sector.

* Channel – This file contains residential and commercial site information, related to meter sensor connection and location information, over a range of years. There are many records in this dataset per metered site. Channel will not be included in the database.
* CI11 to CI41, CI61 – These 11 files contain characteristics specific to commercial sites. With the exception of CI11 that has one record per site, the other nine files have many records per site and is difficult to link to the hourly site level meter data.
* CI51 to CI58 - These 10 datasets do not include site number. Instead, they contain various code numbers that map to code numbers contained in other datasets to provide additional information. We have found most of the datasets they map to, summarized in Table 5 below:

Table 5. Commercial Codes Mapping

|  |  |  |  |
| --- | --- | --- | --- |
| Code Number Source File Name | Characteristics File Name | Characteristics File Variable Name | Description |
| CI51 | ? | ? | ? |
| CI52 | CI31 | CI31005 | Use\_Code |
| CI53 | ? | ? | ? |
| CI54 | CI32 | CI32021 | Equipment\_Code |
| CI55 | CI32 | CI32014 | Control\_Code |
| CI56 | CI32 | CI32010 | Primary\_Fuel |
| CI56 | CI32 | CI32011 | Secondary\_Fuel |
| CI57 | CI41 | CI41006 | Dist\_System\_Type |
| CI58 | CI32 | CI32006 | System\_Type |

* ELCAP88 - This file contains survey data administered in 1988 to residential sites. All the data are numeric. The data are unique by site ID, so they can easily be linked to the hourly site level meter data.
* ENDUSES / LOGHIST– These two files contain information on all building and weather metered sites. They have several fields in common, including site number, di\_integer, begin\_date, end\_date, vcode, and three fields for meter messages. However, they seem to be tracking different data. They are not unique by site ID, so it is difficult to link them to the hourly site level meter data.
* FRC – This file contains one record per residential site and is easily linked to the hourly site level meter data. It has numerous fields with useful residential building characteristics that may be utilized as dimensions in the cube database structure.
* HR## - There are 23 datasets that do not include any site numbers from the meter data and are not included in the database. The 23 files begin with the prefix “HR” followed by a two-digit number and contain survey information.
* NEGCHAN – This file does not appear to contain useful residential characteristics information, but is included in the database. It has only five fields: di\_integer, begin\_date, end\_date, channel\_list, and site number. Since it is not unique by site ID, it is difficult to link this file to the hourly site level meter data.
* PARTICIP – This file contains a mapping of which sites participated in the various surveys conducted, including ELCAP87, ELCAP88, PNWRES83, ROS86, ROSM85, and ROST85.
* PSEUDO – This file contains information for residential, commercial, and weather meter sites. It appears to have a mapping of site numbers to pseudo site numbers. It is not included in the database.
* PX – This file contains only five fields (begin date, end date, site number, di number, and px number). DI number and site number are almost always equal. All 499 metered sites appear in this file. It is not unique by site ID, so it is difficult to link it to hourly site level meter data.
* RI10 to RI81 (except RI72) – These files contain additional residential site characteristics. It is unclear if some of this information also appears in the Charac file. All 24 files have three fields in common: site number, key, and flag. Most are not unique by site ID so it is difficult to link this file to hourly site level meter data.
* RI72 - This dataset does not include site number. Instead, it contains code numbers that map to code numbers contained in other residential datasets to provide additional information.
* ROS86 – This file contains survey data from 1986 for residential sites. All the data are numeric, and the file is unique by site ID.
* SAME – This file contains information on residential sites. There are only six fields on the file, and they are all labeled “SAME” with a year range suffix such as “8385.” The file is unique by site ID.
* VERIFY – This file contains transactional data on metering of residential and commercial sites. It does not contain building characteristics. It is not unique by site ID, so it is difficult to link this file to hourly site level meter data.
* UTILBILL / UTILRATE – The UTILBILL file contains billing data for residential and commercial sites from 1981 to 1989. The UTILRATE file does not contain site numbers. It contains a mapping of utility ID number to utility name, sector, usage type, and rate. The rates are old and therefore are not included in the database.

## HRCP.DAILY.PADS.XPT and HRCP.WEEKLY.PADS.XPT

These two SAS transport files contain one dataset each. The datasets include the same fields, except one is at a daily level and the other is aggregated to a weekly level. Since the two files are both aggregated hourly data and do not contain site numbers for which meter data exists, they are not included in the database.

## NWPP.SASDB.XPT

The four datasets in the NWPP.SASDB.XPT file contain hourly system loads for Bonneville Power Administration, and regulated, private, and public utilities in the Pacific Northwest. The hourly loads range from January 1, 1974 to December 31, 1987. These four datasets cannot be joined with the hourly site level meter data, so they are not included in the database.

## USRD.SASLIB.XPT

The datasets in this SAS transport file all contain the site number field, so they can be mapped back to the hourly site level meter data. These datasets contain additional building characteristic information such as utility billing data, appliance counts, occupancy demographics, and home size, for residential sites.

* Bill83 contains a utility ID field, showing the site number to utility ID relationship. However, the relationship between utility ID and site number is not consistent from file to file, making it difficult to develop this relationship. Given that the information is from 1983, it is not included in the database.
* CRMatrix / CURC / One / Renamed – These files contain the same number of sites and mostly the same columns (see Appendix H). With the exception of One, they also contain the same number of observations. One is a summarization of the CURC file so that it contains one record per site number; therefore, CURC is not included in the database. Renamed is also excluded from the database since it is a duplicate of CRMatrix. One is unique by site ID so it can easily be linked to the hourly site level meter data. CRMatrix is not unique by site ID and cannot be linked to the hourly site level meter data.
* ELCAP86 / ELCAP87– These files appear to contain survey data administered in 1986 and 1987 to residential sites. The 1986 file does not contain a date dimension; the 1987 does (survey date). All the data in both files are numeric. There is a data quality issue with the ELCAP87 file. It is supposed to be a file with one record per site number, but site 90 appears in the file twice. There is also a file named ELCAP88 that is in a different SAS transport file. It contains different fields than the ELCAP86 and ELCAP87 files, so they cannot be appended to one another. Both files are unique by site number and can easily be linked to the hourly site level meter data.
* RI / RIOne – These two files appear to be the same, as they have the same number of unique site numbers and total observations. They also have the same fields (see Appendix I). The RIOne file is not included in the database. RI is unique by site number and can easily be linked to the hourly site level meter data.
* ROST85 / ROSM85 / PNWRES83 – These three files contain survey data from 1983 and 1985. The fields on each file vary, as does the number of sites and observations, so they cannot be appended to each other. The files are unique by site number and can easily be linked to the hourly site level meter data.
* SMatrix and RMatrix – As Appendix J shows, these two files have the same columns. The record counts appearing in Appendix E make it look likely that the files contain duplicate information so RMatrix is not included in the database. SMatrix is not unique by site ID and cannot easily be linked to the hourly site level meter data.
* SCodes and SFlags contain residential characteristics such as met site codes, ZIP code, and climate zone. The files are unique by site ID and can easily be linked to the hourly site level meter data.

### PNNL AVAX Files

Most of the AVAX files appear to be duplicates of the PNNL datasets extracted from the SAS transport files, described above. The AVAX files are saved in various text formats that are not easily transferable to a flat file format for importing into a relational database. Appendix K provides a list of the AVAX files, the estimated number of observations on each file, an indicator of whether the file is already in a flat file format, and comments on the file contents. Appendix L provides a list of the variable names on each AVAX file. For files that appear to be duplicates of PNNL datasets, Cadmus will use the PNNL dataset and ignore the AVAX file.

One AVAX file has an important field for linking the weather stations to the site numbers in the hourly site level meter data. The Metsites file has a field labeled Closest NWS, which is populated with weather station identification numbers. The file also contains a site number field, which can be used to link the file to the hourly site level meter data.

## Summarized Data

The summary files are stored as either text files or Microsoft Excel workbooks, grouped into six categories: commercial, HELM, industrial, irrigation, residential, and RSDP. These files provide hourly site level meter data condensed into average load shapes, coincident factors, and load factors by month, week, day type, and end use, and provide end-use intensities for various building types. They are included in the database and are available as downloads through the Web portal.

## Reports

Numerous Adobe Acrobat and Microsoft Word document reports are included with the ELCAP data. These are listed in Table 6, along with a brief description of the reports contents. These reports provide the results of analysis conducted on the ELCAP data and background information on the project. They are included in the database and are available as downloads through the Web portal.

Table 6. Reports with Description of Contents

| File Name | Description |
| --- | --- |
| Analysis of Load Shape Factors.doc | Brief documentation of the methodology, data, assumptions, and preliminary values calculated for Load Shape Factors (LSF) used in long-term load forecasting tool, based on ELCAP data. |
| Commercial Load Shapes (2).pdf | “The End-Use Load and Consumer Assessment Program: Characterization of Commercial Load Shapes by Weather Day Type.” Prepared for Bonneville Power Administration, by Pacific Northwest Laboratory |
| ELCAP Data Quality Summary.docx | Explanation of why total loads may not match the sum of individual end-use loads for a site, by Rob Pratt at PNNL |
| ELCAP Lighting and Equipment.pdf | “Comparison of ELCAP Data with Lighting and Equipment Load Levels and Profiles Assumed in Regional Models.” Prepared for Bonneville Power Administration, by Pacific Northwest Laboratory |
| ELCAP\_Analysis.pdf | “Significant ELCAP Analysis Results: Summary Report.” Prepared for Bonneville Power Administration, by Pacific Northwest Laboratory |
| ELCAPReportsBibliography.pdf | A bibliography of reports pertaining to or utilizing ELCAP data, from 1984 to 1991 |
| ELCAPReportsBibliography-for-RTF.pdf | Same document as bibliography in previous row, except that some reports are highlighted in yellow. |
| ElectricEnergyUseSingleFamily.pdf | “Description of Electric Energy Use in Single-Family Residences in the Pacific Northwest.” Prepared for Bonneville Power Administration, by Pacific Northwest Laboratory |
| End-Use Load and Consumer Assessment Program.pdf | Utility billing records history, procedure, and summary |
| Enduses.pdf | Description of end uses on hourly site level meter files |
| From Logger to Archive.pdf | Explanation of the collection, processing, and verification of ELCAP load data |
| Getting Started.pdf | User guide for analysis of ELCAP data |
| NPPC.docx | Provides an in-depth review of how residential sites and end uses were selected for the ELCAP study, and which time periods contain the most reliable data |
| Problem Identification.pdf | Discusses the PIRCAR system used to track problems in the ELCAP data |
| SUMMARY REPORT – Assignment of Meteorological Stations to ELCAP Residences.pdf | Describes procedures used in the assignment of meteorological stations to each of the ELCAP residences |

# Database Structure and Schema

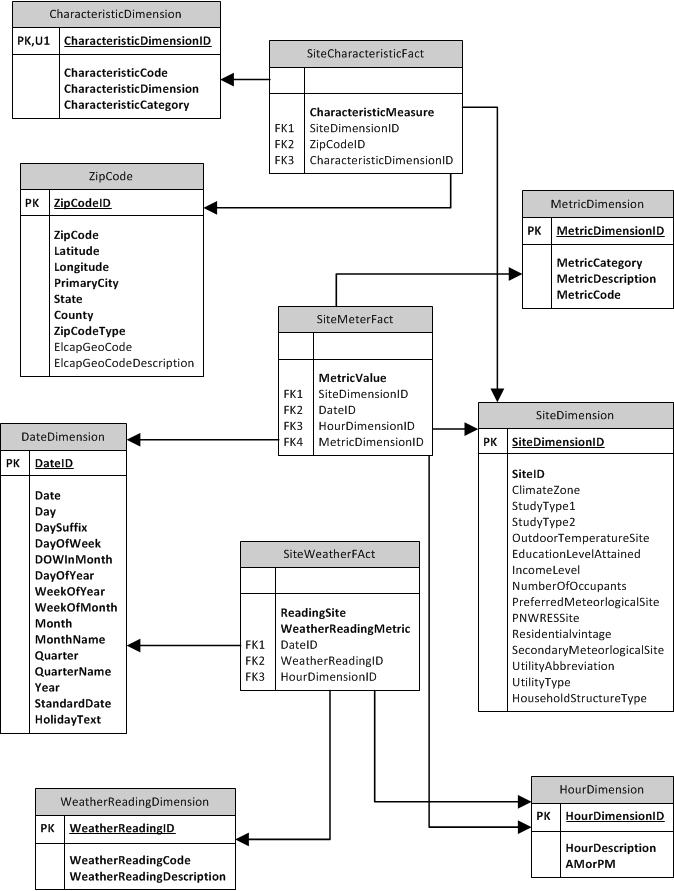
Cadmus recommends that the database be stored in Microsoft SQL Server. Given that the ELCAP database tables contain millions of records, the size of the database is a critical factor in choosing a platform. Microsoft Access, for example, has a two-gigabyte limit on the total size of a database, making it insufficient to handle the ELCAP database. RDBMS (relational database management system) platforms such as Microsoft SQL Server do not have such limitations. The advantage of using Microsoft SQL Server is that it is compatible with Microsoft Access. Microsoft Access can connect to the database through ODBC (open database connectivity) through linked tables, allowing users to write queries and build reports using the external data hosted on a central server.

We designed the structure and database schema diagram shown in Figure 4. The advantage of this multidimensional structure is it is compatible with cube analysis in SQL. Cubes provide an easy-to-use mechanism for querying data stored in a relational database, with quick and uniform response times. Users interact with the cube through an interface that allows them to easily manipulate the controls to obtain customized reports.

The primary tables in this schema are the fact tables: SiteMeterFact, SiteWeatherFact, and SiteCharacteristicFact. These tables contain the values (facts) for each possible combination of variables. Identification numbers in the other columns of the table provide the means for linking to the other tables in the schema, called dimensions. The dimensions provide additional information for each fact. An added advantage of this format is that dimension tables can include extraneous information that is useful for reporting on the facts, such as classifying dates into quarters or ZIP codes into counties.

For example, the SiteMeterFact table contains the hourly meter data for each site in the study. The variable SiteDimensionID links to the SiteDimension table, which provides building characteristics about each site. The variable DateID links to the DateDimension table, which provides numerous categories for grouping data by date, such as quarter, day of the week, and week of the year. The MetricDimensionID links to the MetricDimension table, which provides the different types of metering conducted for each site, such as air conditioning, water heating, and space heating. We added a metric category to this table so that groups of similar end uses such as appliances can be grouped together in reports.

Figure 4. Proposed Database Schema



# Appendices

Due to the large size of the tables in the appendices, we have included them in a Microsoft Excel workbook that accompanies this report.

* Appendix A: ELCAP Hourly Meter Data File Descriptions
* Appendix B: ELCAP Hourly Weather Data File Descriptions
* Appendix C: ELCAP Hourly Site Data Metric Abbreviations Translated
* Appendix D: ELCAP Hourly Weather Data Metric Abbreviations Translated
* Appendix E: PNNL Site Characteristics File Descriptions
* Appendix F: PNNL Site Characteristics File Contents
* Appendix G: PNNL – Comparison of Fields in RESMO Files
* Appendix H: PNNL – Comparison of Fields in ONE, CURC, CMATRIX, and RENAMED Files
* Appendix I: PNNL – Comparison of Fields in RI and RIONE Files
* Appendix J: PNNL – Comparison of Fields in RMATRIX and SMATRIX Files
* Appendix K: PNNL AVAX File Names and Descriptions
* Appendix L: PNNL AVAX File Contents
* Appendix M: Matrix of Site Numbers on PNNL Files

1. Originally, there were 499 total files and 458 hourly meter files. However, 18 commercial sites, four purchase of energy savings sites, and twelve multifamily sites, and one residential site cannot be linked to site characteristics files because they are either pseudo files or they are not included in the characteristics files. In an effort to increase the sample size of the original study, pseudo sites were created for commercial buildings that had more than one business within the overall shell. These sites are not included in the database but will be available to download from the web application. In addition, six weather station files cannot be linked to the hourly meter files and are not included in the database. [↑](#footnote-ref-1)
2. In an effort to increase the sample size of the original study, pseudo sites were created for commercial and multifamily buildings that had more than one business within the overall shell. [↑](#footnote-ref-2)