volta
RESEARCH

Development Of A Data Schema To Hold Energy-Attributes For Residential Housing Archetypes

Final Report: March 2022

Prepared for:
Alex Ferguson
Housing Team Project Lead
Natural Resources Canada
March 31st, 2022

Contents

Introduction

Background

Natural Resources Canada (NRCan) delivers energy labelling programs for residential housing in Canada. These programs include offerings for both new and existing homes. These programs include on-site audits, in which an energy advisor (EA) undertakes numerous measurements of a home's energy-efficiency characteristics. Auditors enter this data into the HOT2000 software, which a) calculates the homes' estimated energy use and b) saves all measurements in an XML formatted file (known as the .h2k file).

Increasingly, NRCan recognized that the data stored in the .h2k files are relevant to various energy conservation and carbon reduction research activities, including stock modelling and mapping and program design. While the data is valuable, its current format is cumbersome. Each house audit exists in a separate .h2k file, and the set of .h2k files comprises various versions and schemas.

This data storage scheme is a barrier to accessing data from NRCan's programs. Researchers and developers must develop their own computer code to parse data from various .h2k formats to perform statistics and analysis. There is no way to search and extract audit data without such measures.

For this reason, NRCan intends to develop an alternate data storage format for information currently contained .h2k file. Key requirements for the new format must include:

- Accommodate data from multiple homes (i.e., multiple .h2k files).
- Data from different versions of .h2k files.
- Be machine-readable and suitable for scripting, machine learning and statistics applications.
- Be interoperable with geospatial analysis applications.

Through this work, NRCan seeks support from data science experts to propose and test suitable data schemas and develop a prototype database containing archetype data provided by NRCan.

Project Details

These are the Project details summarized from the Statement of Work provided to Volta Research by NRCan and reproduced here for the reader's convenience.

Objectives

The objectives of the project are to:

- Recommend an appropriate schema for storing data from arbitrary numbers of .h2k files,
- Build a prototype database using .h2k files provided by NRCan,
- Provide functional scripts demonstrating simple analysis tasks with the prototype database.

These objectives will be investigated through the Scope, Tasks to accomplish the Scope and the project's deliverables.

Scope

The Scope of this project, per the Statement of Work, was limited to:

- **Data contained in the HOT2000 (.h2k file) file:** The schema shall be populated entirely from data within the .h2k file; the current study will not attempt to link records to other databases or populate fields in the schema using data from other sources
- **Measurements collected or inferred by auditors on-site:** These include the physical characteristics of the building such as airtightness, wall and window dimensions, assembly details or effective insulation values. Energy estimates calculated by HOT2000 (and saved in the .h2k file) are to be excluded from the schema
- **Files from recent versions of HOT2000:** The contract will consider only newer, XML formatted data files produced by recent versions of HOT2000 (version 11 or greater). Older, binary file formats (known as .hse files) shall not be in Scope.

Tasks

The following Tasks are given in the Statement of Work and were completed as part of the project. The findings of these tasks follow in the report contents. Each section of the report reflects the content of each task.

- Task 1: Propose format and schemas for housing archetype database
- Task 2: Analyze the proposed schema for portability, scalability, accuracy, and interoperability
- Task 3: Build a prototype database
- Task 4: Develop example analysis scripts
- Task 5: Document schema

Deliverables

The project's deliverables are presented in the report content that follows and the accompanying scripts, databases, and notebook. As part of the Statement of Work, Volta commits to provide the following:

1. A proposed schema for storing HOT2000 data
2. A prototype database populated with the content of 6,800 HOT2000 files provided by NRCan
3. A Jupyter Notebook containing python scripts that provide example analysis using the prototype database
4. A report documenting the proposed schema's fields and relationships to HOT2000 file contents.

The following reporting contents, organized by Tasks, pertain to the first and fourth deliverables, with the remaining deliverables accompanying this report package.

Task 1: Propose format and schemas for housing archetype database

Description

Volta will review the HOT2000 file specification and propose an appropriate schema containing data from multiple home audits. This schema might be based on existing standards for buildings, such as gbXML or some other schema proposed by Volta. The proposed schema will include at least three variants; these will reflect different approaches to reducing model complexity by simplifying geometric details and/or mechanical system definitions.

Findings

Data Schema Design Notes

The data schema proposed in this work is based on a relational database structure, a table-oriented database in which rows represent records (house), and columns represent attributes. If the database is of reasonable size, then the data can be saved in CSV format and easily shared. This row/column data format is easily understandable by many potential users. It does not require a familiarity with object-oriented or NoSQL databases, which can act as a barrier to entry for new users who wish to investigate building science problems but are not familiar with more complex database structures.

This work developed five proposed data schemas that increase in complexity. They are titled:

- Dominant Components CSV
- Equivalent Components CSV
- Supplementary Heating Systems CSV
- Expanded Windows CSV
- Expanded Basements CSV

Each of these formats shares many common features. An accompanying Excel sheet lists all the column names for each of these proposed formats, including descriptions of all attributes. Differences in each of the formats are described below.

Dominant Components CSV

This format aims to consolidate components of the same type into a single representation. For example, all walls will be condensed into a single wall area/height/perimeter, with a single R-Value. This minimizes the number of columns required to represent the house but can result in information loss. The Dominant Components format stores information for each envelope component type based on the most common in the .h2k file. Table 1 shows how an R-Value would be chosen to represent the dominant component by area.

Table 1. Dominant R-Value determination example

Component	Area (m^2)	R-Value (m^2·K/W)
Wall 1	100	6
Wall 2	60	4
Wall 3	50	4
Wall 4	80	3

From this example, the favourite R-Value would be 4 RSI because the total area of walls with this R-Value is 110 m^2, the largest area of walls with the same structure.

This method of selecting a single thermal resistance to represent all components has a crucial flaw, illustrated in the example above - an R-Value may be chosen that does not represent the majority of components (by area). In this example, the most common R-Value makes up 38% of the house's actual walls but would represent all walls in the Dominant Components format.

However, the benefit of this representation is that construction details of the actual home could be easily included in the schema. By basing the overall R-Value on what is present in the house, key building materials can be included that could otherwise lose relevance by calculating an overall effective R-Value.

<u>Equivalent Components CSV</u>

An alternative method of representing all components of a given type in a simple manner involves representing thermal resistance information based on an "equivalent" component. In this case, an effective R-value would be calculated based on parallel heat flow as follows:

$$\frac{A_{tot}}{R_{tot}} = \frac{A_1}{R_1} + \frac{A_2}{R_2} + \cdots + \frac{A_n}{R_n}$$

Where R_{tot} is the total thermal resistance (m^2·K/W) that would be used to represent an entire component type (e.g. walls, basement walls, exposed floors), and A_{tot} (m^2) is the total surface area occupied by that component type.

The benefit to representing the thermal resistance of components in this manner is that, in theory, it should result in equivalent heat flows before and after file conversion. However, using an effective R-value can make it challenging to relate thermal resistance to component construction materials.

It should be noted that effective values for windows cannot be calculated in the same manner since windows are defined by more parameters (e.g., SHGC). As such, both approaches will use the "dominant" approach for windows.

Supplementary Heating Systems CSV

Both previous versions of the CSV format include space for a single supplementary heating system. However, each .h2k file can contain up to five supplementary heating systems. The previous two formats also neglected to include details from other sections of the .h2k file to reduce the number of columns. This extension on the CSV format includes information on the following:

- Five supplementary heating systems (up from one in the previous formats)
- The Additional Openings section of HOT2000
- The Radiant Heating section of HOT2000
- The Multiple Systems section of HOT2000 (in the case of MURBs)

Expanded Windows CSV

This format strives to include more information on windows. It catalogues the five most common windows in a home and tracks the area and count of each window type on each face of the house (based on cardinal direction).

The downside to this format is that if a house file contains less than five types of windows, its row will have several empty columns. However, windows are one of the most important features of a home when determining energy performance, so a solution that more effectively captures window details is required. A sensitivity analysis has not yet been conducted to determine if there is a point of diminishing returns regarding the number of window types to track (i.e., if more or less than five types of windows should be catalogued to optimize accuracy versus unused columns).

Expanded Basements CSV

Basement components are some of the most complex in HOT2000, allowing users to define wall insulation configuration in several ways. An attempt was made to better track basement information by breaking the columns up into many components to better reflect insulated and uninsulated portions of basement walls.

When their performance was analyzed, each of the CSV schemas listed above was built on the previous. That is, the Expanded Windows CSV analysis included information contained within the Supplementary Heating System CSV, and the Expanded Basements CSV included all of the Expanded Windows and Supplementary Heating System CSV information.

Task 2: Analyze proposed schema for portability, scalability, accuracy, and interoperability

Description

Volta will examine the schema (including the proposed variants) in the following subtasks based on:

- **Task 2.1 (Utility):** Is the schema suitable for analysis and application development? What tools will analysts and developers need to use to access the data?
- **Task 2.2 (Scalability):** Can the schema contain an arbitrary number of homes (1,000 — 1,000,000)?
- **Task 2.3 (Portability):** Will the size of the resulting database be suitable for distribution on the web and analysis on cloud or desktop computing hardware?
- **Task 2.4 (Interoperability):** Does the schema adhere to a recognized standard? Do the schema fields correspond to those commonly found in similar databases (e.g., Housing property records?) Do their records use similar terminology and vocabulary?
- **Task 2.5 (Accuracy):** Do simplified variants provide acceptable resolution of building characteristics? Do they offer advantages with respect to scalability or portability?

Findings

Task 2.1: Utility

As part of ensuring that the proposed data format is both suitable for various analysis tools and has broad applicability and utility to a large audience, a non-exhaustive scan of commonly used data analysis, energy simulation, building information file formats, and BIM/architecture software tools with relevance to Part 9 of the National Building Code (NBC) was performed to assess their compatibility with the proposed file format, if applicable. Consultations were also completed with energy and social science academia, geographical information systems professionals, and industry stakeholders using building-related data.

Data Analysis Software

Globally, Excel is estimated to have around 800 million users[1]. In contrast, Python is estimated to have 8.2 million users worldwide as of Q4 2019[2] and, among data science-focused software tools, has the largest percentage of userbase at over 65%[3]. The CSV format from Task 1 was chosen for its compatibility with various software packages. Table 2 shows that the most used data analytics and machine learning platforms (including Excel) natively support imports and analysis of CSV file formats. Though other compatible data formats exist (such as JSON or XML), the CSV format provides a very intuitive column format that is also human-readable

[1] https://medium.grid.is/excel-vs-google-sheets-usage-nature-and-numbers-9dfa5d1cadbd

[2] https://slashdata-website-cms.s3.amazonaws.com/sample_reports/ZAamt00SbUZKwB9j.pdf

[3] https://www.kdnuggets.com/2020/06/data-science-tools-popularity-animated.html

before being loaded into data analysis software and which can be converted to other file formats if required.

Table 2. Top Analytics/Data Science/ML Software[4]

Software	2019 % share	2018 % share	2017 % share	Can import CSV
Python	**65.8%**	65.6%	59.0%	Yes
RapidMiner	**51.2%**	52.7%	31.9%	Yes
R Language	**46.6%**	48.5%	56.6%	Yes
Excel	**34.8%**	39.1%	31.5%	Yes
Anaconda	**33.9%**	33.4%	24.3%	Yes
SQL Language	**32.8%**	39.6%	39.2%	Yes
Tensorflow	**31.7%**	29.9%	22.7%	Yes
Keras	**26.6%**	22.2%	10.7%	Yes
scikit-learn	**25.5%**	24.4%	21.9%	Yes
Tableau	**22.1%**	26.4%	21.8%	Yes
Apache Spark	**21.0%**	21.5%	25.5%	Yes

Energy Simulation Software

Though the data in this proposed CSV format was derived from approximately 6,800 HOT2000 XML-based files supplied by NRCan that are also used for the Housing Technology Assessment Platform (HTAP), it is important to consider its relative usage with other energy simulation software. Though this list is not exhaustive, it highlights some of the potential candidate simulation engines targeted at similar use cases as HOT2000 that could generate energy outputs over a variety of time scales for the given building characteristics input data contained in the proposed format. The energy simulation data for several different simulation engines could eventually be stored in a companion database(s) in the proposed format. This would allow a comparison of simulation energy outputs for a given input data set. This project does not consider the actual translation tools used to take the information within the proposed format. Nonetheless, simulation engines listed in Table 3 indicate if the input file format has the potential to be a candidate for file translation tool development to perform the task of

[4] Ibid.

assembling the respective input file from the proposed CSV file format. Like the building information file formats and BIM/architectural software sections that follow, the input files for each building energy modelling software listed in Table 3 can also be used to derive further inputs to the proposed CSV data format if those are needed in the future. It is also important to note that ASHRAE 140 compliant software can be used to meet NBC 9.36.5 and has the potential to expand the common NBC Part 9 building energy modelling software used in Canada beyond HOT2000. Therefore, it is also noted in Table 3 whether each of the respective building energy modelling software packages is ASHRAE 140 compliant.

Table 3. Building Energy Modelling Software

Software	Software Maintainer (country)	Potential energy output/database input use case (open readability)?	ASHRAE-140 compliant?	Website
Passive House Planning Package	Passive House Institute (Germany)	Yes (xlsx/xlsm format)	Yes (ASHRAE 140-2017)	PHPP – the energy balance and Passive House planning tool
Home Energy Scoring Tool (HES)	Department of Energy (USA)	Yes (API interface or HPXML import/translator)	No	https://hescore-documentation.labworks.org/home-energy-scoring-tool
RESNET HERS Index	RESNET (USA)	Non-standard input format, rating defined by a standard	Yes (ASHRAE 140-2017)	Mortgage Industry National Home Energy Rating Systems Standards
Weatherization Assistant	Oak Ridge National Laboratory (USA)	Yes (.mdb and .csv custom file formats)	No	SOFTWARE DESCRIPTION – Weatherization and Intergovernmental Programs Support
EnergyPlus	National Renewable Energy Laboratory (USA)	Yes (text-based or GBXML)	Yes (ASHRAE 140-2017)	EnergyPlus
ESP-r	The University of Strathclyde (Scotland)	Yes (text-based)	Yes (ASHRAE 140-2017)	ESP-r | University of Strathclyde
eQUEST	James J. Hirsch & Associates (USA)	Yes (text-based)	Yes (ASHRAE 140-2017)	eQUEST
APACHE	Integrated Environmental Solutions Limited (Scotland)	Yes (GBXML)	Yes (ASHRAE 140-2017)	Apache | IES
TRNSYS	Thermal Energy System Specialists (USA)	Yes (text-based)	Yes (ASHRAE 140-2017)	TRNSYS

Interoperable Building Information File formats

More general than the software-specific file formats, several working groups intend to create more generic building information exchange formats. These are similar to HOT2000 files before modelling but differ in how they define their "language" used to refer to their components, granularity of component definitions, and the type and volume of data included beyond the building components. The file formats also may or may not be associated with other activities, including databases, such as HPXML with Home Energy Score (HES) and BuildingSync with Building Energy Data Exchange Specification (BEDES). One of the takeaways from this portion of the project investigation is that schema builders/translator scripts should be created that allows the proposed format to be used to create various building energy modelling input files. A summary of identified file formats is shown in Table 4.

Table 4. Building Information File Formats

File Format Name	Maintainer (country)	Supported software list	Potential integration?	Link to Schema/Notes
Green Building XML (gbXML)	Green Building XML Schema Inc (USA)	gbXML Green Building - Supported Software	Yes (schema terminology translator needed)	gbXML Green Building - Current Schema
Home Performance eXtensible Markup Language (HPXML)	NREL (USA)	openStudio (through NREL's OpenStudio-HPXML)	Yes (schema terminology translator needed)	HPXML Data Dictionary v3.0.0 Standardized (BPI-2200-S-2013) and a strong tie in with HES
Industry Foundation Classes (IFC)	BuildingSMART (USA)	Certified Software - buildingSMART International	Yes (schema terminology translator needed)	Standardized (IFC Schema Specifications - buildingSMART Technical)
BuildingSync	Alliance for Sustainable Energy, LLC (USA)	Use Cases	Yes (schema terminology translator needed)	Schema Viewer built on BEDES

BIM/Architecture Software

Architectural software's building energy model file outputs are important potential inputs for the proposed CSV format. What follows is a selection of common BIM/Architecture software packages:

- Autodesk's Revit
- Graphisoft's ARCHICAD
- Trimble's SketchUp
- Tekla's BIMsight software
- Adobe Acrobat, FME Desktop
- Constructivity Model Viewer
- CYPECAD

Of particular interest is that usage of Revit has increased for Part 9 building types in recent years. This was determined through consultation with Canadian energy advisors indicating that it was increasingly common to receive a Revit file rather than a drawing to work off of for significant residential retrofits or building projects. ARCHICAD is also another popular software package in global markets. Both software packages can import/export building energy modelling compatible files to/from gbXML or directly to/from PHPP. Future translation tools that can take advantage of ingesting/outputting gbXML in addition to .h2k files into/from the proposed file format could provide a significant pathway to increase industry participation in the expansion of the file format.

Stakeholder Consultations

We performed limited stakeholder consultation as part of the project activities to ensure that potential users find value in the proposed file format. We consulted with building and social science academia members involved in single-family housing research. From the academic stakeholders, we generally heard that CSV was the best format because it was straightforward to use for all levels of researchers. We also bread the following specific feedback:

- Directionality of glazing would be nice to have for shading and other directional considerations (e.g., trees, neighbouring buildings),
- HOT2000 and EnergyPlus are single zone models using lumped elements in elevations, so average per elevation wall thermal resistance would be sufficient,
- The number of individual windows and doors could help, wall area, attic area, etc. to assist with costing research activities,
- The most critical need is to tie in with the geographic database to use the data meaningfully.

We also consulted with the energy advisor industry and software developers that operate in the building energy space in Canada. The feedback from software developers generally is that as long as they can load the data into a database and information about the fields is available, they can work with various formats. A CSV format will serve them as well as a JSON or other format, and the end-use specifics depend on the business problem they are solving. The CSV format was best for most of the EA firms spoken to, although few of them found that they would immediately be able to act on the data other than those already thinking about this and participating in other data or energy modelling initiatives. EA firms also noted the increasing usage of Revit for the building information that is submitted to them for new and high-end existing homes.

Finally, we obtained feedback from federal and municipal government stakeholders to understand how they might leverage the data format. Some specific feedback follows:

- The output of this would have relevance to the national building layer and should be considered in that context
- There would be value in participating and sharing this with the Open Geospatial Consortium and integrating it with their building information interoperability work
- Municipalities like the idea of being able to open a CSV to view basic information on housing that they can aggregate, but they need the geo-references to do so

- That tie ins with demographic are important to making program decisions, and that should be something that the format enables

Task 2.2: Scalability

Each row in the CSV format produced as part of this task holds the information from one house. Given the potential tools and software explored in Task 2.1 that could be used to perform analyses with the proposed CSV format, Microsoft Excel is one of the most popular and important to consider for the proposed format. However, Excel has the most limitations in its maximum acceptable file size (dimensions) of the tools investigated. In this case, file contents are limited by rows and columns in Excel, but rows are the limiting factor for the CSV file format proposed in this project. Therefore, around 1 million rows can be represented in the proposed format – 1 million homes – without exceeding Excel's basic data loading capability. This provides even the novice data user with a significant opportunity to analyze the proposed format without using more advanced software. The 1,000,000 rows also represent almost the entire size of the current EnerGuide database – of which many files are not in the .h2k file format – and therefore provide good value in terms of scalability and software accessibility for EnerGuide data.

Task 2.3: Portability

The data size for a 6,800-house file is less than 10 MB, a very compact download size. For a file of 1,000,000 houses, we can expect an uncompressed file size of around 1.4 GB. Even with 1,000,000 homes, data sizes are reasonable given current internet services and can be reduced for distribution with compression. For comparison to other Government of Canada data, StatsCan Census of the Population data downloads would be of comparable sizes, with some exceeding 2 GB[5].

As mentioned in Task 2.1, the file is easy to load in Excel and other tools once downloaded. The data can be used by simply filtering in Excel and, for more complex work, easily loaded in Python (or other data science applications). Therefore, we can expect in most cases the proposed data format will be portable and has a low barrier to entry.

Task 2.4: Interoperability

Though the proposed format and project focused on interoperability with the .h2k file format first, the project investigated several other similar building characteristics databases globally to assess the potential to model this proposed format on another database or align it to allow interoperability. This investigation determined that alignment is certainly possible in future interactions of the database format, and the format has the potential to foster international collaboration on interoperability.

A common issue found while investigating other residential information databases was that they either focused primarily or entirely on performance data or contained limited information on building characteristics and construction. However, there are many promising candidates for

[5] https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E

database harmonization. They are presented in Table 5 with abridged information, including comments on their interoperability potential and links to the most current schemas for each, if available.

Recommendations for future work involve joining working groups and standards committees to look at the harmonization of either this proposed format with the schema's, categorical, and numerical contents of similar databases. These groups were investigated during this project, but not exhaustively, and potential was found in joining Building Energy Performance ASHRAE Technical Committee 7.6 or engaging with International Building Performance Simulation Association, but others may also be better suited or more applicable

Table 5. Candidates for Database Harmonization

Database Name	**Description from database maintainer (Comments from Volta, if applicable)**	**Country of Origin**	**Website and schema links (if available)**
ResStock Analysis Tool	Researchers at the National Renewable Energy Laboratory (NREL) developed ResStock. It provides large-scale residential energy analysis by combining: ● Large public and private data sources ● Statistical sampling ● Detailed sub-hourly building simulations ● High-performance computing. (This is an interesting tool because it shows what the proposed data format could be used for within a Canadian context. The proposed format could potentially feed into the data sources used by this platform.)	USA	ResStock US Building Stock Characterization Study
National Residential Efficiency Measures Database	The National Residential Efficiency Measures Database is a publicly available, centralized resource of residential building retrofit measures and costs for the U.S. building industry. With support from the U.S. Department of Energy, NREL developed this tool to help users determine the most cost-effective retrofit measures for improving the energy efficiency of existing homes. (The data contained within this database is very close to what we might see in a HOT2000 file and is, therefore, a good candidate for interoperability)	USA	National Residential Efficiency Measures Database
The EU Building Stock Observatory (BSO)	The EU Building Stock Observatory (BSO) was established in 2016 as part of the clean energy for all Europeans package and aimed to provide a better understanding of the energy performance of the building sector through reliable, consistent, and comparable data. (This database contains a very broad set of data, but the residential construction components data is somewhat limited compared to some of the USA databases. There is potential for interoperability and collaboration. Some similar headings and components are used as in the proposed format)	EU	EU Building Stock Observatory EU Buildings Database | energy
Residential Energy Consumption Survey (RECS)	EIA administers the Residential Energy Consumption Survey (RECS) to a nationally representative sample of housing units. RECS collected data from more than 5,600 households in housing units statistically selected to represent the 118.2 million housing units occupied as a primary residences. Data from the 2015 RECS are tabulated by geography and for particular characteristics, such as housing unit type and income, that are of particular interest to energy analysis. The results of each RECS include data tables, a microdata file, and a series of reports.	USA	Residential Energy Consumption Survey (RECS) Residential Energy

	(Very similar to StatsCan housing dimensions data and a good source for a large, statistically relevant dataset. Some opportunity to look at the naming interoperability of building components)		Consumption Survey (RECS) - Data - US Energy Information Administration (EIA)
Building Performance Database (BPD)	The Building Performance Database (BPD) contains information about the building's energy use, location, and physical and operational characteristics. The BPD compiles data from various data sources, converts it into a standard format, cleanses and quality checks the data and provides users with access to the data to maintain anonymity for data providers. The BPD consists of the database itself, a graphical user interface allowing data exploration, and an application programming interface allowing the development of third-party applications using the data. (The database is well documented with API access, but the component granularity is limited. The proposed CSV format contains many more fields than BPD, and Interoperability potential is possible but will need further investigation.)	USA	Building Performance Database API Fields - BPD API Documentation
Municipal Property Assessment Corporation (MPAC)	MPAC enables the purchase of Assessment, Site, Structural, and Sales data on all types of properties across Ontario individually or in bulk. Information is refreshed from the database weekly to ensure users have access to the most current property information in Ontario. (MPAC database is behind a paywall but contains validated data on real estate and municipal tax roll centric components such as the number of bedrooms and area of floors. It is similar to what can be obtained through StatsCan data. There is limited potential for interoperability, but it is worth investigating for initiatives like the national building layer.)	Ontario, Canada	MPAC propertyline™ >> Products & Services >> Product Catalogue & Pricing
Building Energy Data Exchange Specification (BEDES)	The Building Energy Data Exchange Specification (BEDES) is a dictionary of terms and definitions commonly used in tools and activities that help stakeholders make energy investment decisions, track building performance, and implement energy-efficient policies and programs. (The database is a relevant initiative, and though very broad and well beyond the residential landscape, alignment with it would piggyback on the BEDES' alignment with other standards. A good candidate for interoperability investigation.)	USA	Bedes Manager
Standard Energy Efficiency Data Platform (SEED) Platform	The Standard Energy Efficiency Data Platform (SEED) Platform is an open-source software application designed to manage building performance data (such as required by a benchmarking ordinance) which can be costly and time-consuming for states, local governments, and other organizations. SEED helps users combine data from multiple sources, clean and validate it, and generate queries and reports.	USA	SEED Platform™ Documentation

	(This would be more considered a data management platform than a collection of data, so the opportunities for integration are more limited. A good example of how the proposed CSV could be leveraged)		
Open Database of Buildings (ODB)	The Open Database of Buildings (ODB) is a collection of open data on buildings, primarily building footprints, and is made available under the Open Government License - Canada. The ODB brings together 65 datasets originating from various government sources of open data. The database aims to enhance access to a harmonized collection of building footprints across Canada. (Useful information is limited to footprint data which could be of some value in the validation of the main floor area within the proposed data format)	Canada	The Open Database of Buildings
StatsCan Census of the Population (Housing specific dimensions)	A detailed statistical portrait of Canada and its people by their demographic, social and economic characteristics. (The housing-specific dimensions in the census offer information such as building type, condition, and size that could be used for interoperability. More importantly, linking the two data sources will allow a much broader sociodemographic analysis of EnerGuide data in the future.)	Canada	Census Profile, 2016 Census Download, Census Profile, 2016 Census

Task 2.5: Accuracy

Data Loss Analysis

File "reconstructors" were built to attempt to quantify potential data loss when converting HOT2000 data into the various proposed CSV schema formats. These python scripts took the CSV schema information and re-built .h2k files based only on the information contained within the schemas. Data loss was then estimated by simulating the reconstructed files and investigating variances in the following simulation outputs:

- Total Annual Energy Consumption, [GJ] (Results, Annual, Consumption, total)
- Design Heating Load, [W] (Results, Other, designHeatLossRate)
- Design Cooling Load, [W] (Results, Other, designCoolLossRate)

The performance of each schema was then assessed based on the percent difference of each of the above variables from that of the source file.

Due to time constraints, simulations were run in HOT2000's General mode for both the source and reconstructed files. Since simulations in EnerGuide (ERS) mode typically take at least 5 seconds, the process of reconstructing and simulating the files would have been very time-intensive. However, simulations can still be run in ERS mode later to compare results under standard operating conditions. Additionally, for simplicity, MURBs and mobile home files were not assessed in this exercise since they have not yet been converted into the latest version of HOT2000 (v11.11). However, their parameters are still included in the proposed schemas.

Simulation Results

Figures 1, 2, and 3 display boxplots that summarize the distributions of percent differences in total annual energy consumption, design heating load, and design cooling load, respectively. Similarly, Table 6, Table 7, and Table 8 include descriptive statistics of percent differences in total annual energy consumption, design heating load, and design cooling load, respectively. These tables also have a measure of the percentage of reconstructed files that are within 5% of the target source file value. This 5% threshold is used as a general indication of performance.

Interestingly, performance between the dominant and equivalent component CSVs was very similar. This is likely because both CSV schemas use the dominant window component, and windows are a key factor in building performance that, in this case, overshadows simplifications made elsewhere. However, this can only be asserted in the context of the analysis provided, which involved a comparison of HOT2000 results. It is unclear whether one method or the other would provide additional value in other use cases. As such, both the dominant and equivalent component definitions have been included in the schema provided with this work.

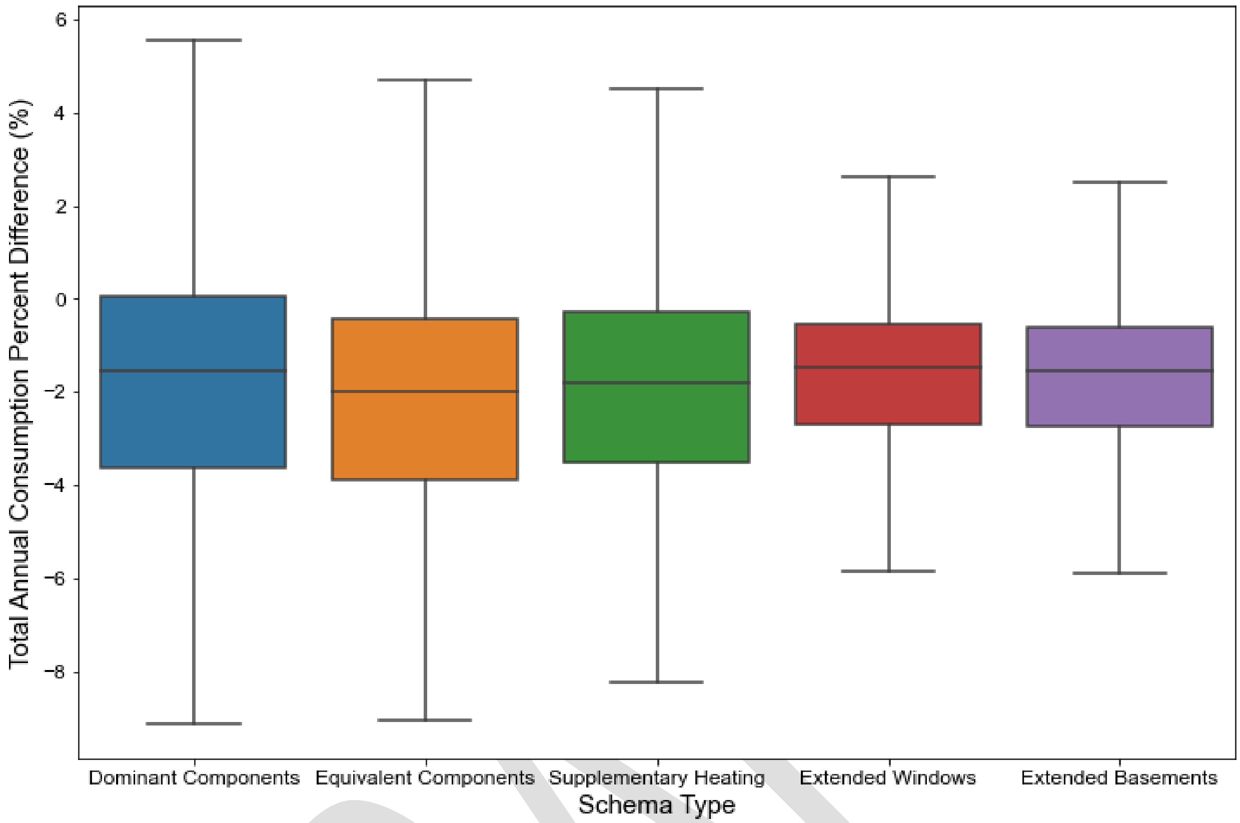

Figure 1. Distribution of total annual consumption percent difference for each proposed schema.

Table 6. Descriptive statistics of total annual consumption percent difference for each proposed schema.

Schema Type	**Dominant Components**	**Equivalent Components**	**Supplementary Heating**	**Extended Windows**	**Extended Basements**
Mean	-1.80	-2.20	-1.84	-1.60	-1.74
Standard Deviation	4.46	3.92	3.61	2.45	2.03
Minimum	-37.84	-35.52	-35.52	-18.25	-16.54
25th Percentile	-3.61	-3.89	-3.51	-2.69	-2.73
Median	-1.54	-1.98	-1.82	-1.49	-1.54
75th Percentile	0.07	-0.45	-0.30	-0.56	-0.63
Maximum	32.89	29.37	29.37	22.70	20.42
< 5% of Target	80.37%	80.52%	84.15%	93.26%	94.36%

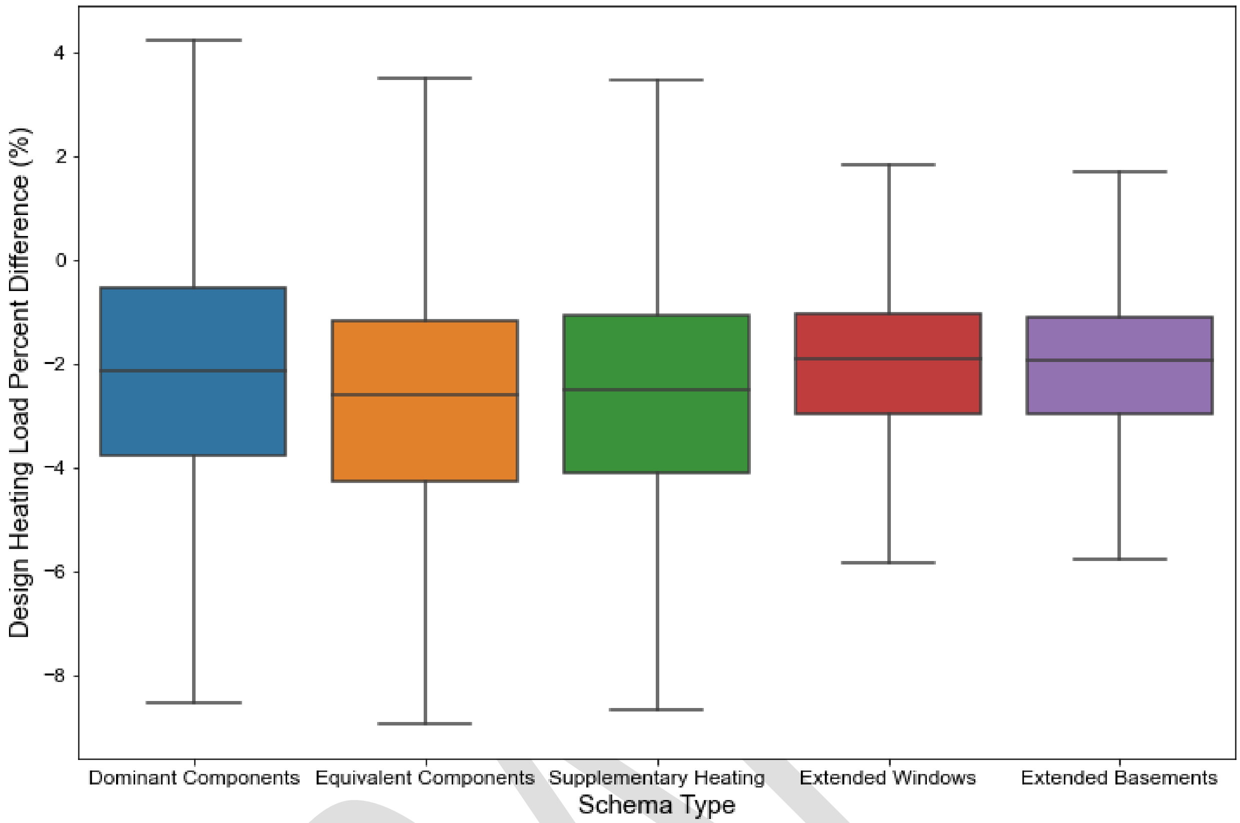

Figure 2. Distribution of design heating load percent difference for each proposed schema.

Table 7. Descriptive statistics of total annual consumption percent difference for each proposed schema.

Schema Type	**Dominant Components**	**Equivalent Components**	**Supplementary Heating**	**Extended Windows**	**Extended Basements**
Mean	-2.13	-2.58	-2.41	-1.80	-1.94
Standard Deviation	4.41	3.71	3.56	2.32	1.89
Minimum	-39.06	-39.13	-39.13	-18.74	-11.68
25th Percentile	-3.75	-4.27	-4.11	-2.96	-2.97
Median	-2.13	-2.61	-2.51	-1.90	-1.94
75th Percentile	-0.54	-1.15	-1.07	-1.03	-1.09
Maximum	39.36	25.57	25.57	25.46	25.46
< 5% of Target	80.69%	79.69%	81.54%	95.44%	96.52%

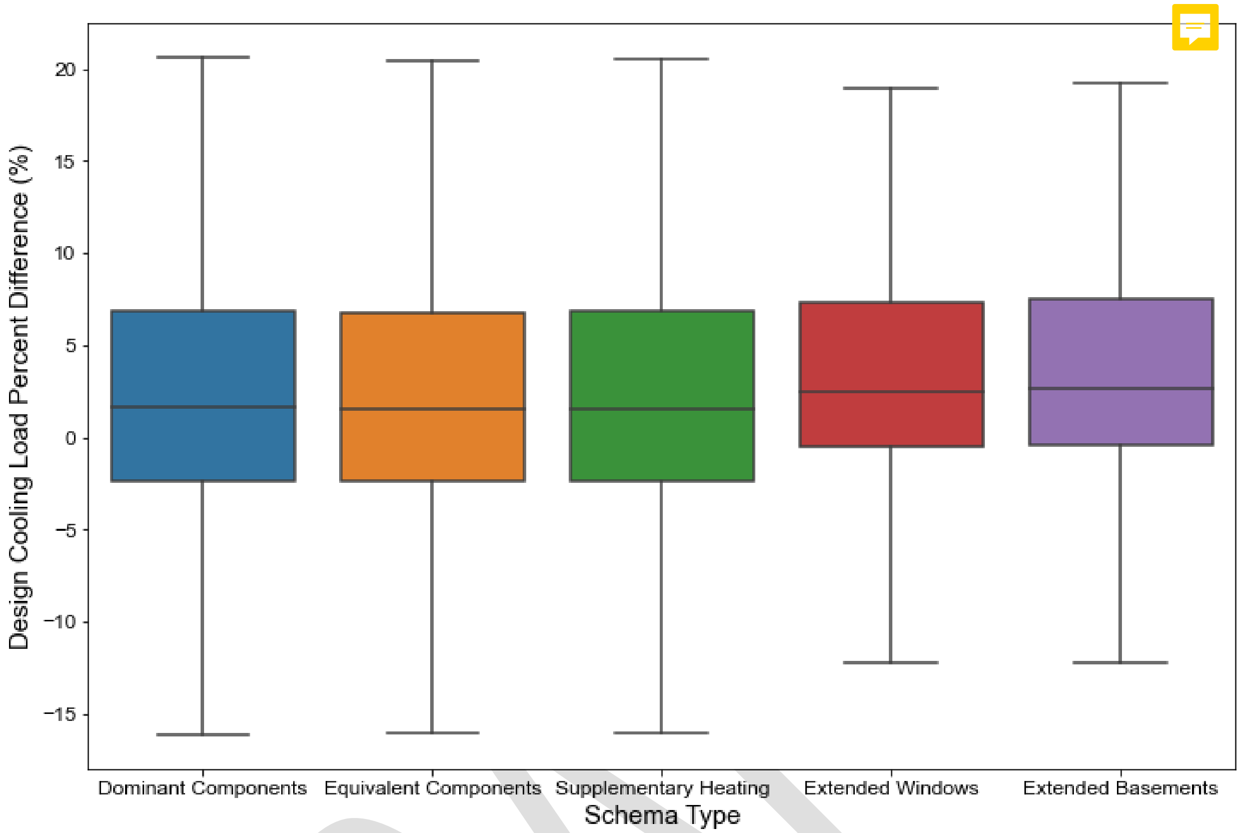

Figure 3. Distribution of design cooling load percent difference for each proposed schema.

Table 8. Descriptive statistics of total annual consumption percent difference for each proposed schema.

Schema Type	Dominant Components	Equivalent Components	Supplementary Heating	Extended Windows	Extended Basements
Mean	2.96	2.85	2.93	4.44	4.62
Standard Deviation	10.29	10.17	10.14	9.50	9.49
Minimum	-72.20	-75.79	-75.79	-75.79	-75.79
25th Percentile	-2.39	-2.41	-2.34	-0.53	-0.42
Median	1.67	1.52	1.58	2.50	2.64
75th Percentile	6.83	6.75	6.81	7.26	7.45
Maximum	198.52	193.80	193.80	283.49	283.49
< 5% of Target	54.49%	55.09%	55.11%	60.69%	60.33%
< 10% of Target	78.27%	79.01%	79.05%	81.49%	81.04%

In terms of total annual consumption, adding the additional supplementary heating system information and expanded window details shows the steepest increase in the number of files within 5% of their targets. While including additional basement details further narrows this range, its improvements are minimal compared to the improvements provided by the additional window details. Similar performance can be observed in design heating load, where the addition of window details comes with the most drastic accuracy increase.

The accuracy of the design cooling load was observed to be more modest than that of the other two key parameters. Again, additional window details showed the largest improvement, but this still only produced 60% of files within 5%. However, further investigation in Table 4 shows that over 80% of files are still within 10% of the target design cooling load. A thorough analysis of the causes of these deviations has not yet been conducted.

This work also found that including additional information on supplementary heating systems and windows led to notable reductions in data loss and improvements in the relative accuracy of simulation results. As such, both expanded schemas have been included in this work. They are presented each in their own CSV file and can be linked to the main schema via file ID.

The Extended Basements format was not found to improve performance by a significant margin and drastically increases the complexity of the proposed schema to capture a small number of edge cases. It is not recommended that this format be further explored, and it has not been provided along with this work. However, it can be made available upon request if needed.

Lessons Learned in File Reconstruction

While this exercise helps to provide insights into potential sources of data loss in the proposed schemas, there are key limitations in the analysis. Firstly, it is limited by the number of compared variables compared, which only highlight the buildings' overall annual and seasonal peak performance of the buildings. Additionally, it is limited by the inputs, outputs, and overall capabilities of HOT2000. Dynamic factors such as hourly loads (other than peak design loads) could not be investigated, nor could changes in complex geometry.

The analysis is also influenced by the file reconstructions, which were developed as a first attempt to achieve the task and have not been perfected. Any user (analyst, engineer, student) could create a different file builder for various simulation engines, each with varying degrees of performance and accuracy. However, the authors hope that the results presented in this work can act as a baseline for schema performance and that future users can expand and improve upon the work presented herein.

Task 3: Build prototype database

Description

NRCan will provide Volta with a set of 6,800 HOT2000 files. Volta will build a prototype housing archetype database based on data imported from these files. As part of this task, Volta will develop a python script that can a) read an arbitrary number of HOT2000 files and b) populate records in the database based on the contents of those files.

Findings

Python scripts were created to read .h2k files, construct a CSV database, and populate it with the 6,800 HOT2000 files per the design in Task 1. The populated CSV databases are provided with this report.

Task 4: Develop example analysis scripts

Description

Volta will author a set of example scripts that perform rudimentary analysis on the prototype archetype database. NRCan intends for these scripts to serve as guides for other academic and industry researchers and developers who will develop their own applications in the future. The functions performed by the example scripts will be agreed to by NRCan and Volta throughout the project; Volta will author the scripts as Jupyter Notebooks.

Findings

Sample analysis scripts have been provided in Jupyter notebook (Python) format. They utilize the pandas and seaborn libraries to analyze and visualize several different aspects of the data. For example, main wall R-Value and air change rate distributions by decade are presented for different provinces. Another example shows the data broken down and visualized to investigate how common each primary heating fuel type is in each province. These are just a few examples of the countless insights that can be gained from this dataset.

Task 5: Document schema

Description

Volta will develop documentation for the proposed schema. This documentation will include a human-readable description of the schema's fields. If the proposed schema includes multiple tables or nested structures, the documentation will include visual entity-relationship diagrams. It will also describe relationships between the database schema and data collected in a HOT2000 data file.

Findings

An accompanying Excel sheet lists all the column names for each of these proposed formats, including descriptions of all attributes.

Conclusions

This work investigated options for data schema formats to contain Canadian residential building archetype data, currently stored in HOT2000 (.h2k, XML) format. An industry scan and literature review were conducted to identify and comment on existing data formats suitable for this application and ways in which the resultant data schema and database could be used.

This work sought to recommend a schema that:

- Can store data from an arbitrary number of .h2k files;
- Can store data from multiple versions of (XML-formatted) HOT2000 files;
- Is machine-readable; and,
- Is interoperable with geospatial analysis applications.

The Scope of the proposed schema only includes *input* data, not simulated energy consumption or other performance data.

It was decided to pursue a data schema modelled after a relational database, as opposed to pursuing an object-oriented structure. Relational databases can be represented as collections of rows and columns and, therefore, can be opened in a wide variety of common analytical tools (e.g. Excel, Python). Additionally, due to the relatively small archetype database size, this structure allows the data to be easily shared in CSV format. This common file format is highly accessible and can be picked up and interpreted by various users with various experience levels.

Provided along with this report are the sample data schema in CSV format, including "extensions," which can be included or excluded depending on the user's needs, and sample scripts in Jupyter notebook (Python) format to demonstrate simple analysis tasks that can be performed with the prototype schema database. A companion Excel sheet has also been provided, including all column header names and descriptions of their content.

Recommendations

From the analyses on data loss, it is recommended the misaligned cooling loads be investigated further and data loss be explored by producing other data formats from the CSV database.

This exercise has opened the door for several future potential investigations, some of which may include:

- Investigating data integrity and performance when converting the proposed schema data for other use cases and simulation engines.
- Investigating the addition of more detailed geometry data, which HOT2000 currently does not track.
- Determining improvements to the schema that can enhance its ability to capture information critical to evaluating peak cooling loads.
- Investigating use cases that can enhance geospatial analyses for Canadian housing.
- Investigating interoperability and international cooperation with organizations in the USA and EU to share data and resources. This participation could happen directly or through participation in standards committees.