PAPER • OPEN ACCESS

The Building Data Genome Directory – An open, comprehensive data sharing platform for building performance research

View the article online for updates and enhancements.

The Building Data Genome Directory – An open, comprehensive data sharing platform for building performance research

Xiaoyu Jin[1], Chun Fu[2], Hussain Kazmi[3], Atilla Balint[3], Ada Canaydin[3], Matias Quintana[2,4], Filip Biljecki[2], Fu Xiao[1], Clayton Miller[2,*]

[1] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong
[2] College of Design and Engineering, National University of Singapore (NUS), Singapore
[3] Department of Electrical Engineering, KU Leuven, Belgium
[4] Future Cities Laboratory Global, Singapore-ETH Centre, Singapore

E-mail: *clayton@nus.edu.sg

Abstract. The building sector plays a crucial role in the worldwide decarbonization effort, accounting for significant portions of energy consumption and environmental effects. However, the scarcity of open data sources is a continuous challenge for built environment researchers and practitioners. Although several efforts have been made to consolidate existing open datasets, no database currently offers a comprehensive collection of building data types with all subcategories and time granularities (e.g., year, month, and sub-hour). This paper presents the Building Data Genome Directory, an open data-sharing platform serving as a one-stop shop for the data necessary for vital categories of building energy research. The data directory is an online portal (`buildingdatadirectory.org/`) that allows filtering and discovering valuable datasets. The directory covers meter, building-level, and aggregated community-level data at the spatial scale and year-to-minute level at the temporal scale. The datasets were consolidated from a comprehensive exploration of sources, including governments, research institutes, and online energy dashboards. The results of this effort include the aggregation of 60 datasets pertaining to building energy ontologies, building energy models, building energy and water data, electric vehicle data, weather data, building information data, text-mining-based research data, image data of buildings, fault detection diagnosis data and occupant data. A crowdsourcing mechanism in the platform allows users to submit datasets they suggest for inclusion by filling out an online form. This directory can fuel research and applications on building energy efficiency, which is an essential step toward addressing the world's energy and environmental challenges.

1. Introduction
The rise of artificial intelligence as a tool for built environment applications has the potential to impact several industries significantly. However, data availability in the built environment domain remains a critical bottleneck due to privacy concerns and acquisition costs [1]. Open data sources are essential for understanding energy consumption patterns, identifying areas for improvement, and testing energy-saving strategies, especially in the absence of in situ measurements. Yet, access to open data sources in the built environment domain lags behind
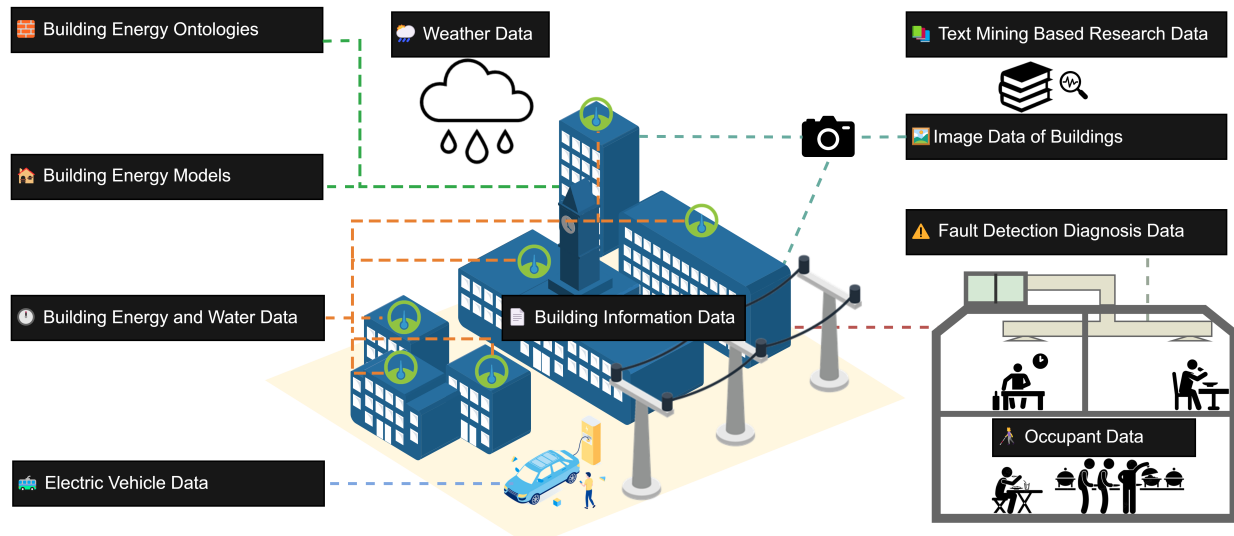
Figure 1. Schematic of the categories of datasets included in Building Data Genome Directory

other communities [2], posing limitations for researchers and practitioners in developing effective energy-saving solutions [3]. In addition to limited accessibility, available open datasets are often dispersed and require labor-intensive and time-consuming collation due to varying formats and sources [4]. Efforts have been made to aggregate open datasets and share them through platforms or directories such as the Building Performance Database (BPD) [5], the Building Data Genome (BDG) projects [6, 7], and the Directory of Buildings Energy Consumption Datasets (DBECD) [8]. However, these projects have limitations in the diversity of data types, lack of user contributions, and missing data.

This paper outlines the development of a comprehensive data-sharing platform for building performance research. This effort is achieved by creating a data directory that is publicly available and includes functions for filtering, visualization, and uploading new data sets. The Building Data Genome Directory is a lightweight web app that links to a wide range of open datasets, offering users easy access to comprehensive coverage of relevant information. In subsequent sections, the paper will introduce the data sources, data category definitions, reasons for inclusion, critical functions of the web app, and some application cases.

2. Data sources

The directory focuses on collecting information about open building performance datasets that are widely dispersed and fragmented, which conventionally would require a rigorous data collection process. Metadata for the directory was gathered from various open data sources, including government disclosure programs, research projects, institutes, and publicly available dashboards. Details on each of these data source categories are discussed in the following subsections. The directory data sources are divided according to category and type of data based on the format (e.g., tabular, image) and process of the system that created the data (e.g., HVAC, occupants, sensors). Figure 1 shows an overview of the data set categories, which will be outlined in the following subsections.

2.1. Government disclosure data

Data from government disclosure programs is a significant source for built environment data. One example is the Local Law 84 (LL84) of New York City (NYC) in the United States,

which requires building owners to disclose their energy and water consumption data through benchmarking annually [9]. This directive has led to the publication of the Energy and Water Data Disclosure dataset for Local Law 84 by the NYC government. These city-level datasets can contain many samples, with some featuring tens of thousands of buildings, although they may have coarse-grained time intervals of a year or a month. To collect these datasets, a comprehensive review of relevant literature and examination of laws pertaining to data disclosure was conducted [1]. Open data portals provided by city governments [10], such as the NYC open data portal (`https://opendata.cityofnewyork.us/`), were also browsed to gather available datasets, ensuring the comprehensiveness of the data directory.

2.2. Open research data
Research institutes and organizations have published various datasets for building performance research. Some datasets are available on websites, such as the Building Data Genome dataset on Kaggle [6] or the 3D city model of Singapore public housing buildings on GitHub [11]. Other datasets are published through journals, with *Scientific Data* being a significant venue. A recent review has also listed open-source datasets for building energy demand [2]. These datasets typically provide detailed information about individual buildings but may not have large numbers of samples (generally less than 5,000). A common differentiator of these types of data sets is that the time-series frequency may be higher, sometimes even at the minute level, offering a more granular view of a building's energy usage. Some datasets also provide detailed information about building characteristics, solar installations [12], morphological indicators [13], or sensor locations and building structure [14]. Accessing and leveraging these datasets allows researchers to gain comprehensive insights into individual buildings and their energy usage. To collect these datasets, relevant reviews and research papers were examined, including platforms that provide access to datasets referenced in articles,

2.3. Data collected from open, online dashboards
In response to the growing emphasis on net-zero and sustainability goals in the higher education sector, many educational institutes and universities, such as the University of California, Berkeley, Cornell University, and Princeton University, have public energy management dashboards that provide access to energy usage data for further study and analysis. For these datasets, a data acquisition pipeline can be built using scripts to automate the process of extraction from these dashboards, enabling batch downloads of performance data from thousands of buildings. The directory includes several datasets that were retrieved from these types of public web-based energy management dashboards. For many of these dashboards, the API of the data source can usually be found using built-in web browser developer tools. Once the data API is identified, an automated process can be configured with the required data parameters, such as building ID and specific time period, to enable batch downloading of performance data from a web-based dashboard.

3. Overview of the directory interface
The Building Data Genome Directory can be found online at: `buildingdatadirectory.org/`. The interface comprises of a main page, referred to as the *Meta Directory*, which provides an overview of all available datasets and several sub-pages presenting datasets by types. The *Meta Directory* page introduces the Building Data Genome Directory and outlines the scope of the collected datasets. As a web app, it has filtering, visualization, and uploading functions for the datasets. Datasets pertaining to buildings, such as *Building Energy and Water* and *Building Information* provide geospatial granularity levels that correspond to individual buildings or, at the very least, communities, instead of the aggregated data of an entire city. The *Meta Directory* includes a schematic diagram showcasing the various datasets available in the Building

Table 1. Categories of the data in the directory with short descriptions and an example representative dataset of each type.

Category Name	Data Scope Description	Representative Dataset
Building Energy Ontologies	Haystack, Brick Schema, and datasets supporting the semantic model development	A Synthetic Building Operation Dataset [15]
Building Energy Models	Building energy model information and simulation data of prototype buildings	Building Stock dataset [16]
Building Energy and Water Data	Time-series energy data and water consumption data of buildings	Energy and Water Data Disclosure for Local Law 84 [9]
Electric Vehicle Data	Charging infrastructure datasets and time-series energy consumption data from charging sessions	Campus Electric Vehicle Charging Stations Behavior
Weather Data	Historical data from environmental sensors and weather prediction	Campus Electric Vehicle Charging Stations Behavior [17]
Building Information Data	Stock characteristics, GIS data, and project management data for a large number of buildings	European Building Stock Characteristics Dataset [18]
Image Data of Buildings	Image data of a large number of buildings	Annotated Image Database of Architecture [19]
Text Mining Based Research Data	Text-mining data from previous research and communities	A Comprehensive Text-mining Driven Review of Scientific Literature [20]
Fault Detection and Diagnosis Data	Ground-truth and simulated datasets for anomalies in the built environment and building systems	Large-scale Energy Anomaly Detection (LEAD) Dataset [21]
Occupant Data	The thermal comfort data of occupants collected from experiments	Cozie smartwatch application [22]

Data Genome Directory, as shown in Figure 1. Each black label in the diagram represents a specific data type and has a corresponding subpage, with its link conveniently located on the left column of the web page. The scope description for these types and the representative datasets are presented in Table 1. The *Add New Dataset* uploading function is at the bottom of the left-hand button. Users must fill in the *Dataset Name*, *URL*, and *Dataset Type* items to submit a possible contribution to the directory. The datasets submitted by the users will be stored and displayed at the bottom of the *Meta Directory* page, and they will be added to the directory after undergoing a review process.

The category with the highest number of data sets is *Building Energy and Water*, which includes over 30 datasets at the moment. A metadata table that provides essential information about the datasets is displayed on this page, including disclosure status (e.g., data opening level, license availability, organization) and information on the building samples. Figure 2 shows the filtering and visualization functions. The filtering functions enable users to select datasets by location, time interval, and building type. The visualization functions include bar plots with adjustable axes to visualize numerical information, bubble plots to display sample and variable numbers with the size of circles denoting sample sizes and variable quantities, and heatmaps to visualize variable categories.

4. Conclusion and future works

The Building Data Genome Directory is a potentially valuable resource for building energy research, providing comprehensive datasets and web app functions for filtering, visualization, and uploading. This directory can be a starting point for researchers and analysts who want to start the exploration process for applicable open data sets for their studies. Numerous research

Figure 2. Building Data Genome Directory interface showcasing the filtering and visualization functions.

endeavors are anticipated to emerge as branches stemming from this directory. As highlighted by Jin et al. [1], the availability of comprehensive datasets will significantly expedite research in building energy, encompassing areas such as building energy management, grid management, and socio-economic analysis. The team is developing a sub-branch within the Building Data Genome Directory focusing on time-series feature analysis utilizing energy consumption data.

4.1. Future expansion and data quality considerations
Future work can optimize the directory by improving functions such as allowing brief dataset descriptions during uploading and incorporating semantic searching capabilities. Enhancing search capabilities for different data types, such as geographic location, would also improve usability, as well as considering unconventional data sources such as scraping relevant data on buildings from property websites [23] and considering volunteered geographic information such as OpenStreetMap [24] in locations that have data of reliable quality. Finally, to strengthen the crowdsourcing aspect of our platform, we plan to implement a functionality to allow users to flag erroneous information and allow trusted users to edit the database. Building a community around the directory would foster user communication and optimize the web app. Collecting feedback and insights through discussions and forums would provide valuable inputs for enhancing features and usability. By actively engaging with users, the directory can continue to evolve and serve as a valuable resource for building energy researchers b

References

[1] X. Jin, C. Zhang, F. Xiao, A. Li, C. Miller, A review and reflection on open datasets of city-level building energy use and their applications, Energy and Buildings 285 (2023) 112911.

[2] H. Kazmi, Í. Munné-Collado, F. Mehmood, T. A. Syed, J. Driesen, Towards data-driven energy communities: A review of open-source datasets, models and tools, Renewable and Sustainable Energy Reviews 148 (2021) 111290.

[3] F. Johari, G. Peronato, P. Sadeghian, X. Zhao, J. Widén, Urban building energy modeling: State of the art and future prospects, Renewable and Sustainable Energy Reviews 128 (2020) 109902.

[4] X. Jin, F. Xiao, C. Zhang, Z. Chen, Semi-supervised learning based framework for urban level building electricity consumption prediction, Applied Energy 328 (2022) 120210.

[5] T. Walter, M. D. Sohn, A regression-based approach to estimating retrofit savings using the building performance database, Appl. Energy 179 (2016) 996–1005.

[6] C. Miller, F. Meggers, The building data genome project: An open, public data set from non-residential building electrical meters, Energy Procedia 122 (2017) 439–444.

[7] C. Miller, A. Kathirgamanathan, B. Picchetti, P. Arjunan, J. Y. Park, Z. Nagy, P. Raftery, B. W. Hobson, Z. Shi, F. Meggers, The building data genome project 2, energy meter data from the ashrae great energy predictor iii competition, Scientific data 7 (1) (2020) 368.

[8] T. Babaei, H. Abdi, C. P. Lim, S. Nahavandi, A study and a directory of energy consumption data sets of buildings, Energy and Buildings 94 (2015) 91–99.

[9] C. E. Kontokosta, Predicting building energy efficiency using new york city benchmarking data, Proceedings of the 2012 ACEEE Summer Study on Energy Efficiency in Buildings, Washington, DC, American Council for an Energy-Efficient Economy (2012).

[10] B. Lei, R. Stouffs, F. Biljecki, Assessing and benchmarking 3D city models, International Journal of Geographical Information Science 37 (2023) 788–809.

[11] F. Biljecki, Exploration of open data in southeast asia to generate 3d building models, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences VI-4/W1-2020 (2020) 37–44.

[12] A. N. Wu, F. Biljecki, Roofpedia: Automatic mapping of green and solar roofs for an open roofscape registry and evaluation of urban sustainability, Landscape and Urban Planning 214 (2021) 104167.

[13] F. Biljecki, Y. S. Chow, Global Building Morphology Indicators, Computers, Environment and Urban Systems 95 (2022) 101809.

[14] N. Luo, Z. Wang, D. Blum, C. Weyandt, N. Bourassa, M. A. Piette, T. Hong, A three-year dataset supporting research on building energy management and occupancy analytics, Scientific Data 9 (1) (2022) 156.

[15] H. Li, Z. Wang, T. Hong, A synthetic building operation dataset, Scientific data 8 (1) (2021) 213.

[16] N. M. Frick, E. Wilson, J. Reyna, A. Parker, E. Present, J. Kim, T. Hong, H. Li, T. Eckman, End-Use load profiles for the U.S. building stock: Market needs, use cases, and data gaps (Nov. 2019).

[17] T. H. Xiaochen Zhang, S. Grijalva, Campus electric vehicle charging stations behavior (2021).

[18] N. Milojevic-Dupont, F. Wagner, F. Nachtigall, J. Hu, G. B. Brüser, M. Zumwald, F. Biljecki, N. Heeren, L. H. Kaack, P.-P. Pichler, et al., Eubucco v0. 1: European building stock characteristics in a common and open database for 200+ million individual buildings, Scientific Data 10 (1) (2023) 147.

[19] J. Chen, R. Stouffs, F. Biljecki, Hierarchical (Multi-Label) Architectural Image Recognition and Classification, in: Proceedings of the 26th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA) 2021, Vol. 1, 2021, pp. 161–170.

[20] M. M. Abdelrahman, S. Zhan, C. Miller, A. Chong, Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature, Energy and Buildings 242 (2021) 110885.

[21] C. Fu, P. Arjunan, C. Miller, Trimming outliers using trees: winning solution of the large-scale energy anomaly detection (LEAD) competition, in: Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 456–461.

[22] P. Jayathissa, M. Quintana, T. Sood, N. Narzarian, C. Miller, Is your clock-face cozie? A smartwatch methodology for the in-situ collection of occupant comfort data, in: CISBAT2019 Climate Resilient Buildings - Energy Efficiency & Renewables in the Digital Era, Lausanne, Switzerland, 2019.

[23] X. Chen, F. Biljecki, Mining real estate ads and property transactions for building and amenity data acquisition, Urban Informatics 1 (2022) 12.

[24] F. Biljecki, Y. S. Chow, K. Lee, Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes, Building and Environment 237 (2023) 110295.