Dataset: An Open Dataset and Collection Tool for BMS Point Labels

Gabe Fierro
UC Berkeley
gtfierro@cs.berkeley.edu

Sriharsha Guduguntla
UC Berkeley
sguduguntla@berkeley.edu

David E. Culler
UC Berkeley
culler@cs.berkeley.edu

ABSTRACT

Semantic metadata standards for buildings such as Brick and Project Haystack show promise in enabling wide-scale deployment of energy-efficiency measures and advanced building management technologies. However, techniques for converting existing diverse and idiosyncratic forms of building metadata to these standard forms is an area of active research. To encourage and facilitate research into the development and evaluation of such techniques, we are releasing an open dataset of metadata pulled from real building management systems, containing attributes for 103,064 points over 92 buildings. In addition, we are releasing an open-source tool for scraping and cleaning metadata from BMS for contribution to the dataset.

ACM Reference format:
Gabe Fierro, Sriharsha Guduguntla, and David E. Culler. 2019. Dataset: An Open Dataset and Collection Tool for BMS Point Labels. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 2 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent efforts to establish semantic metadata standards for buildings span academia (Brick [3]), industry (Project Haystack [1]) and standards bodies such as ASHRAE (223P [2]) and W3C (BOT [7]). These standards are promising avenues for consistent descriptions of buildings and their cyber-physical resources, ultimately facilitating the development and deployment of energy-efficiency measures and advancements in building management and operation at scale. However, the deployment of these standards is hampered by the manual effort required to normalize existing, heterogeneous building descriptions to a given standard. For these standards to experience wide-adoption, they must provide tooling to normalize existing building descriptions.

Existing digital representations of buildings have many forms ranging from non-structured human-readable annotations to extensive industrial standards. One prominent source of such building metadata is the monitoring and control networks contained in large commercial buildings. These networks are commonly accessed through building management systems (BMS) and supervisory control and data acquisition systems (SCADA) which present a digital interface to the sensors, actuators, alarms, statuses and control points present in the building. While the names of these points – termed *point labels* – may contain some semantic information such as the point's name, location, function or related equipment, such labels are often unstructured, building-specific, inconsistent, and reliant upon vendor-specific conventions for consistent interpretation.

```
AHU.AH04.CCO
AHU.AH04.CCV
AHU.AH04.Cooling Enable
AHU.AH04.ECM
AHU.AH04.EF11.Start Stop
AHU.AH04.HCO
AHU.AH04.HCV
AHU.AH04.Heating Enable
```

Figure 1: Point labels for a sample building (single attribute)

```
Eighth Floor/E8126 Machine Room CRAC-9/% Capacity,% Capacity
Eighth Floor/E8126 Machine Room CRAC-9/Fan Run Hours,Fan Run Hours
Eighth Floor/E8126 Machine Room CRAC-9/Humidifier Run Hours,Humidifier Run Hours
```

Figure 2: Multiple point attributes including engineering units for a sample building

Recent work has demonstrated success in inferring structured semantic models like Brick from unstructured point labels, employing techniques such as human-in-the-loop for learning interpretations of point labels from domain experts [5], transfer learning for generalizing parsing rules from one building to another [6], and combining lexical clustering of point labels with timeseries analysis [4]. However, it is difficult for these techniques to demonstrate robust performance due to a lack of access to large amounts of diverse building point label data. The application of machine learning methods to the task of metadata normalization is also hampered by a lack of data.

Collecting building point label data is difficult because such metadata is sequestered behind corporate firewalls, distributed between proprietary systems which must be accessed using number of industrial protocols such as BACnet, and often exposes sensitive information such as the names of buildings or rooms. To address these issues and assist in the building metadata standardization effort, we are releasing: (a) a dataset of BMS point labels from real buildings, and (b) a tool for generating and cleaning point label dumps from building BMS.

2 POINT LABEL DATASET

The dataset consists of a set of attributes for each point in a building management system, distributed as CSV files (one per building). Most buildings in this first release contain a single attribute ("point

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

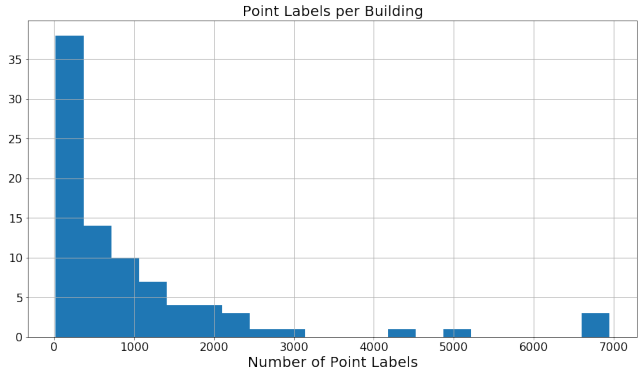

Figure 3: Distribution of the number of BMS point labels available per building in the released dataset.

Generic Operation	Before	After
Strip *n* chars from left or right	`NAE 05 N2 2 VMA 126 ACTCLGSP`	`ACTCLGSP`
Find and replace substrings	`A2.S2HALL:DMPR COMD`	`A2.S2HALL:DPR COMD`
Select field by position with given delimiter	`AHU.01.CCV`	`CCV`
Extract matches using regular expressions	`SODC1SP03__FLT`	`FLT`

Table 1: Examples of string operations supported by the web interface, applied to point labels

label") per point (Figure 1). A few buildings contain multiple attributes in addition to point labels (Figure 2). We plan to expand the dataset with other metadata carried by building management systems such as engineering units and textual descriptions where available.

At time of writing, the dataset contains attributes for 103,064 points pulled from the BMS of 92 buildings. This includes point dumps for 5 buildings from the Brick reference models referenced in [3]. Figure 3 contains the distribution of the number of BMS point labels available per building: the average number of point labels is 1120 and the standard deviation is 1,640.

We have applied some simple preprocessing to the dataset to remove identifying information revealing such as names of buildings and rooms, but have otherwise preserved the content and structure of the point attributes. As a result, the quality of the data mirrors the idiosyncratic and fragmented nature of real building metadata. Accounting for this state of affairs is an important task for automated metadata normalization techniques and tools.

3 DATA COLLECTION TOOL

In addition to the dataset, we are releasing a tool to substantially reduce the effort in extracting point labels and related metadata from BMS and preparing this data for public release. The tool begins by scanning a network for BMS endpoints which it can connect to; the tool then pulls all available points and metadata from the BMS. The tool then organizes this compiled data into a CSV format and loads it into a web interface which is presented to the user. The user, such as a building manager, can then visualize, clean and prepare the dataset for public release. Our tool provides a set of string operators for editing undesired information out of the building metadata. Table 1 contains examples of several of these operators, which include field extraction, regular expression matching and find/replace. To deal with outliers, the tool also offers the ability to edit individual labels within the web interface.

After cleaning and filtered the data, the user can use the tool to publish the finished dataset and integrate it with the dataset we are releasing with the paper. As the dataset expands over time, it will contain a higher number and a more diverse population of building metadata, giving researchers a rich body of data from which to develop metadata normalization methods.

4 FUTURE WORK AND CONCLUSION

We have currently only implemented a BACnet/IP adapter for the data collection tools, but we plan to expand the set of adapters to include other common BMS technologies such as KNX, OPC and LonTalk. The tool is available at https://github.com/gtfierro/point-label-sharing, and will be packaged for general use soon. Determining a suitable online repository is also a subject of future work. The initial release of the dataset will be available as a zipped collection of CSVs, but we would like to have live storage that can be added to over time.

This work presents an open dataset of building point attributes for use in developing and evaluating data-driven metadata normalization methods for buildings. To enable the maintenance and growth of the dataset, we are also releasing an open-source tool for collecting point attributes from BMS and applying simple string cleaning rules for removing private and identifying information in anticipation of publicly releasing the data. We hope that this first step encourages the donation of point attribute data from the community to facilitate research into metadata normalization and promote the use of standardized semantic metadata models in buildings.

REFERENCES

[1] 2018. Project Haystack. http://project-haystack.org/.
[2] American Society of Heating, Refrigerating and Air-Conditioning Engineers. 2018. ASHRAE's BACnet Committee, Project Haystack and Brick Schema Collaborating to Provide Unified Data Semantic Modeling Solution. http://web.archive.org/web/20181223045430/https://www.ashrae.org/about/news/2018/ashrae-s-bacnet-committee-project-haystack-and-brick-schema-collaborating-to-provide-unified-data-semantic-modeling-solution.
[3] Bharathan Balaji, Arka Bhattacharya, Gabriel Fierro, Jingkun Gao, Joshua Gluck, Dezhi Hong, Aslak Johansen, Jason Koh, Joern Ploennigs, Yuvraj Agarwal, et al. 2016. Brick: Towards a unified metadata schema for buildings. In *Proceedings of the ACM International Conference on Embedded Systems for Energy-Efficient Built Environments (BuildSys). ACM.*
[4] Bharathan Balaji, Chetan Verma, Balakrishnan Narayanaswamy, and Yuvraj Agarwal. 2015. Zodiac: Organizing Large Deployment of Sensors to Create Reusable Applications for Buildings. ACM, 13–22.
[5] Arka A Bhattacharya, Dezhi Hong, David Culler, Jorge Ortiz, Kamin Whitehouse, and Eugene Wu. 2015. Automated metadata construction to support portable building applications. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments.* ACM, 3–12.
[6] Jason Koh, Bharathan Balaji, Dhiman Sengupta, Julian McAuley, Rajesh Gupta, and Yuvraj Agarwal. 2018. Scrabble: transferrable semi-automated semantic metadata normalization using intermediate representation. In *Proceedings of the 5th Conference on Systems for Built Environments.* ACM, 11–20.
[7] Mads Holten Rasmussen, Pieter Pauwels, Christian Anker Hviid, and Jan Karlshøj. 2017. Proposing a central AEC ontology that allows for domain specific extensions. In *2017 Lean and Computing in Construction Congress.*