2020 Building Performance Analysis Conference and
SimBuild co-organized by ASHRAE and IBPSA-USA

EXTRACT USEFUL INFORMATION FROM BUILDING PERMITS DATA TO PROFILE A CITY'S BUILDING RETROFIT HISTORY

Wanni Zhang[1], Tianzhen Hong[1], and Xuan Luo[1]
Lawrence Berkeley National Laoratory, Berkeley, CA

ABSTRACT

Building retrofit is one of the key strategies for cities to reduce energy use and GHG emissions. The historical information about changes to buildings is crucial to infer the buildings' current energy system efficiency levels and to identify candidate buildings for retrofit. In general, a building permit is required before the start of any construction activity of a building, such as changing building structure, remodeling, or installing new equipment. Moreover, many large cities provide public datasets of building permits in history. Therefore, the permits are a potentially good resource for mining information on the city's retrofit history. In this study, we use the permit dataset from the city of San Francisco as a case study. Location and time information from the dataset is also used to depict the retrofit timeline of each building and the whole building stock. The type of work of the permit is inferred from the descriptive text by a machine learning model. At last, the limitations of the current permit dataset and potential improvements on the permit data management are discussed to better utilize the information in the future.

INTRODUCTION

In the United States, building permits are required before the start of any construction activity to enlarge, repair, add to or demolish a building, to renovate the structure and to change equipment or systems in buildings, etc. Generally, the issued permit records are maintained by the corresponding city agency (usually the Department of Buildings), and many cities provide digitalized records of building permits for public access, which contain a lot of information including but not limited to the issued date, current status, building type, geographical information of the building and a short descriptive text of the scope of work. All the information together depicts the changing history of the city's building block. This can be used to infer the retrofit frequency, equipment lifetime and the current energy efficiency of each building, which are all helpful for urban scale building simulations.

However, in our knowledge, limited studies are found to explore the building permits data. Some studies focus on developing tools for the display of permits dataset to benefit citizens and public authorities (Vert & Vasiu, 2017) (Eirinaki et al., 2018). Only basic analysis is done with the existing features from the dataset, such as filtering different types of permits. Another study (Stevenson, Emrich, Mitchell, & Cutter, 2010) uses the clustering method on the date and location information of the permits to assess spatial and temporal trends in recovery activities following the hurricane. Some study uses text mining on building maintenance work orders, which has a similar nature with permit records, but the dataset is limited to a small group of buildings (Gunay, Shen, & Yang, 2019). Generally, the major challenge lies in handling large-scale text data, which contains the most information about the retrofit work. This study uses the text mining approaches to extract useful information from the building permit datasets over the past several decades, to bring insights on building retrofit and urban planning. The rest of the paper first describes the methodology of the whole process, including data collection, data preparation and text mining. The machine learning models used for text mining is briefly explained. Then the results of the data analysis from San Francisco dataset are illustrated. Finally, limitations and future work are discussed.

METHODOLOGY

The study uses the permits data from San Francisco (SF) as a case study. The timeline analysis is first conducted on the whole building block, and the permits for each building are identified to analyze the activities of an individual building.

The historical permits of San Francisco are tabulated for download and keep updated periodically on the open data portal of the city. The most updated datasets by the beginning of the study (October 2018) are used for analysis. All the permits are separated into four datasets: Building Permits, Electrical Permits, Boiler Permits and Plumbing Permits. The Boiler Permits dataset is excluded since it is all about the permission for operating boilers, which is not relevant to our scope of the study. Based on the description provided by the data source and San Francisco Department of Building Inspection (DBI), who issued most of the permits, the Building Permits dataset covers a broad scope of work, including the construction renovation, the interior change and the mechanical work, etc. and the

Plumbing Permits dataset is for all types of plumbing and mechanical permits. Hence the rest of the three datasets are combined for analysis.

The data fields of the tabulated dataset include the time and the status of the permit, the permit category (new construction, demolition or alteration of existing construction), the usage and location of the building, and a short descriptive text about the scope of work, which is written in natural language by different people. Some examples are shown as follows.

- *32nd fl ti includes the demo of non-structural partitions and building new non strctural partitions. all new or revised. hvac ele, plumb, ls will be engineered. mep under separate permit*
- *at roof level, construct new mechanical room, 194 sf, with new interior stair from 2nd floor. new rear 1 level deck.*

It can be observed that the sentences are not usually well-structured, and there are often abbreviations and typos in the text. Secondly, although the word "hvac" appears in the first permit, the last sentence indicates that mechanical, electrical and plumbing ("mep") work is actually not included in this permit. In the second permit, the appearance of the word "mechanical" doesn't imply the permit covers mechanical work, since it is all about the mechanical room. These suggest that it is hard to understand the content of the permits by simply identifying keywords, since the context of the keywords matters. Therefore, machine learning techniques related to natural language processing (NLP) are introduced to infer the type of work for each record. Details of the machine learning model are described later.

Data Preparation

To classify the type of work of each record, the machine learning model has to be trained with some data labeled data. Since it is unpractical to manually label a large amount of data in SF dataset for training due to the limited resource of human labor, some public-accessible permits datasets that have the categorization of the scope of work are collected from other U.S. cities as the training data. Then a small sample of human-labeled permits from SF dataset is used to test the actual performance of the model. However, the schema of other city's permit dataset varies largely. The categories used in each dataset are listed below. To align the way in which they categorize their permit type, some categories are merged and the intersections of the categories are used, which are Mechanical, Electrical, Building and Plumbing (NYC data only keeps the mechanical and plumbing records).

- Austin: Mechanical, electrical, building, plumbing, Driveway/ Sidewalks
- Philadelphia: Mechanical, Electrical, Plumbing, Alteration, Zoning, Addition, New Construction, Demolition, Other
- San Diego: Mechanical, electrical, building, plumbing, demolition, renewable, other
- NYC: Mechanical, Plumbing, Boiler, Fuel Burning, Fuel Storage, Standpipe, Sprinkler, Fire Alarm, Equipment, Fire Suppression, Curb Cut

It worth mentioning that because one permit can file for multiple types of retrofit work. Therefore, the labels for the training dataset are "multi-hot" encoded as "is_mechanical", "is_building", "is_electrical", and "is_plumbing".

All the collected datasets are cleaned before analysis. Records with required fields missing are removed, and records in SF dataset with the same permit number, current status are considered as duplicated and are removed as well. Besides, the permits are in different status from "issued", "complete" to "expired" and "canceled", but only those completed permits indicate that the work described was indeed implemented, so permits that are completed and have the information for completion date are filtered for analysis, which accounts for around 61% of the whole SF data.

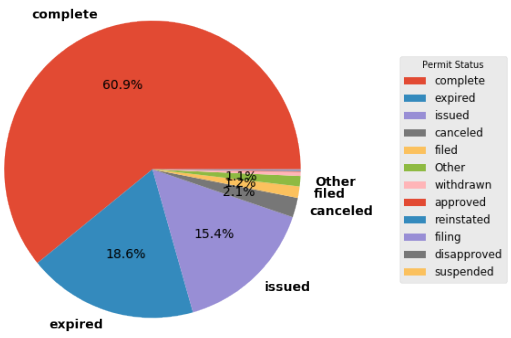


Figure 1 The distribution of permit status in SF dataset

After data cleaning, the useful records of each dataset are summarized as follows:

- San Francisco Building Permits: 501,578 (54.1% of all) completed permits in 1978-2018
- San Francisco Plumbing Permits: 129,413 (68.4% of all) completed permits in 2007-2018
- San Francisco Electrical Permit: 190,407 (81.4% of all) completed permits in 2001-2018
- Austin: 1,087,599 useful records in 1980-2018

- Philadelphia: 523,313 useful records in 2007-2018
- San Diego: 98,660 useful records in 2010-2018
- NYC: 536,989 useful records in 2000-2018

Text Mining

To understand the general types of work covered in each permit, a deep learning model is used to do text mining, which is formulated as a multi-label classification problem. The useful data from Philadelphia, Austin, NYC and San Diego is mixed together as the training dataset, with the assumption that these datasets share similar language patterns with the SF dataset. The output of the machine learning model indicates that whether a permit contains mechanical work, plumbing work, electrical work or building work. After training, 350 randomly-selected permits in the SF dataset are manually labeled as the test set. The model that has the highest test accuracy will be used to infer the type of work of all the SF permits.

Different types of models are tested for the text classification problem. The first one is to use word-embedding with neural network. Word-embedding is a method for feature extraction of the text data, which represents similar words with vectors that have occupied close spatial positions. In this study, we use the 100-dimension GloVe (Global Vectors for Word Representation) (Pennington, Socher, & Manning, 2014) word vectors pre-trained with large Wikipedia corpus as the initial word-embedding to get a good start, and train the embedding parameters along with the process of training the machine learning model. To embed the words from the permit description, the text data is first preprocessed, including lowercasing, tokenizing, removing the useless tokens and stemming. Figure 2 illustrates the process of pre-processing and word-embedding of a permit.

To be fed to the neural network, the inputs need to have the same sequence length. After tokenized, 95% the permit texts have a length of less than 50 tokens, so all the texts longer than that will be truncated and the sequence shorter than 50 will be filled with a placeholder. A convolutional neural network (CNN) modeled after the dynamic convolutional neural network (DCNN) proposed in another study (Kalchbrenner, Grefenstette, & Blunsom, 2014) is used for classification.

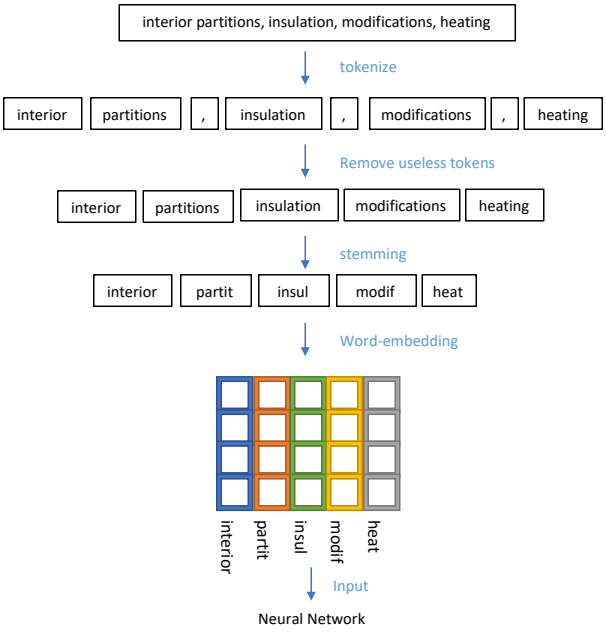


Figure 2 Diagram of data preprocessing and word-embedding process

The other model used for the text classification is the BERT (Bidirectional Encoder Representations from Transformers) published by Google that produced the state-of-the-art results in many NLP tasks (Devlin, Chang, Lee, & Toutanova, 2018). BERT is a language model based on the attention mechanism that provides embeddings conditioned on both the left and right contexts. The attention mechanism is a way to assign different relevance weights to the input sequence to make the model focus on the most relevant context based on what is being processed now. BERT is pre-trained on the huge corpus and applicable to various downstream NLP tasks. For our problem, we add a classification layer on top and finetune all the parameters end-to-end.

In each model, the hyperparameters, such as the learning rate, training epochs are all tuned to achieve the best performance.

Based on the classification results, records with the same "Block", "Lot" and "Street Number" attributes are grouped to calculate intervals of each type of permits for individual buildings, which infers the frequency of each type of retrofit work.

RESULTS

Text Classification

With the aforementioned mixed dataset from Philadelphia, Austin, NYC, and San Diego, 80% is used for training and 20% is for validation. The results of the best performance of the two types of models are listed in Table 1. The validation accuracy (val_acc) and the test accuracy (test_acc) are the accuracy evaluated on the validation data and labeled SF data respectively. The default accuracy is the accuracy achieved by labeling all samples as the majority of the category, which indicates the accuracy of assigning labels with no knowledge of the data.

The performance of the BERT model is overall better than the CNN model with word-embedding, especially on the permits for building work (including structural, interior changes, etc.). The test accuracy of all the types is improved by more than 13%, among which the plumbing type has the highest accuracy. The cause might be that the scope of plumbing work is not that diverse and hence the language features can be more easily captured.

Table 1 Classification accuracy for the type of work of different models

	BUILDING	**ELECTRICAL**	**MECHANICAL**	**PLUMBING**
val_acc_cnn	0.9130	0.8916	0.8744	0.9214
val_acc_bert	0.9231	0.9045	0.9104	0.9490
default val_acc	0.8493	0.7834	0.7489	0.6274
test_acc_cnn	0.6886	0.7057	0.7229	0.8371
test_acc_bert	0.7875	0.7125	0.7475	0.8406
default test_acc	0.5657	0.6257	0.5371	0.6400

Overall analysis

First, some overall information about the whole building block is profiled. As shown in Figure 3, over 77% of the permits are filed for residential buildings, and within the commercial building permits, office and retail make up the majority.

With the selected classification model mentioned above, a total of 821,398 permits are labeled for the type of work. The number of Building permits is the largest among the four, which includes work of interior remodeling, envelope retrofit, structural change, and so on. Less than 10% of the permits contain mechanical work (Figure 4).

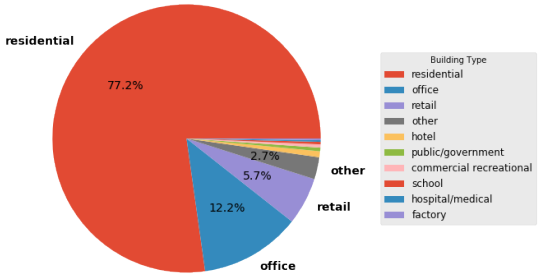


Figure 3 Permits by building type in SF dataset

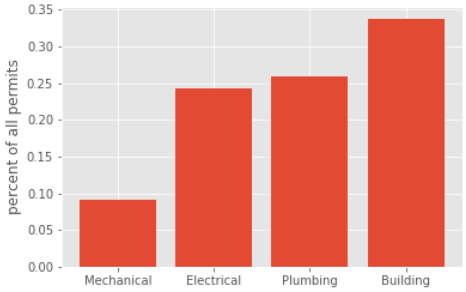
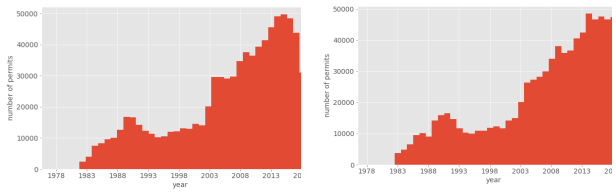


Figure 4 Permits by inferred type of work in SF dataset

The timeline of the permits being issued and being completed are illustrated in Figure 5. The patterns of the two charts generally align with each other, which implies that the permits are usually completed quickly (within one or two years) once issued. The number of permits rose in 2003 and continuously increased after that. This might be partially due to the increased number of buildings and the digitalization of the permit data. Further analysis is needed to see whether this ties to the economic cycle.



(a) *(b)*

Figure 5 Histogram of permits (a) issued date, (b) completed date

Individual building analysis

With the aforementioned method to group permits for each building, there are 173,901 buildings identified in total (after removing the records without location information), which is closed to the size of the whole building stock in SF. As shown in Figure 6 and Table 2, on average, each building has 4.71 permits in history and 97% of the buildings have less than 15 permits in history. This accuracy of the statistics might be affected by the unavailable older permit records or unreported retrofit/renovation work.

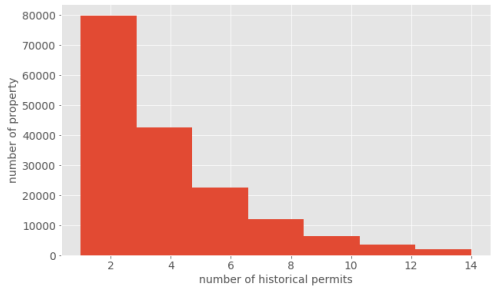


Figure 6 Distribution of the number of completed permits in history for each building

Table 2 Statistics of the number of completed permits in history for each building

MIN	MAX	MEAN	MEDIAN
1	2032	4.71	3

Then the time intervals between permits of each building are analyzed to infer the average cycle of each building having a retrofit. Intervals are calculated by the completed date and the permits completed within less than 30 days are assumed as one single renovation. Figures 7 and Figure 8 illustrate the distribution of the intervals (in years) of all permits and of each type of permits, and Table 3 and Table 4 show the statistics of the intervals respectively.

Table 3 Statistics of time intervals between permits for each building

MEAN	MEDIAN	MAX
4.91	2.05	35.46

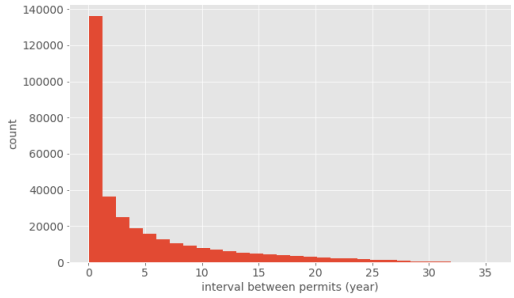


Figure 7 Distribution of time intervals between permits for each building

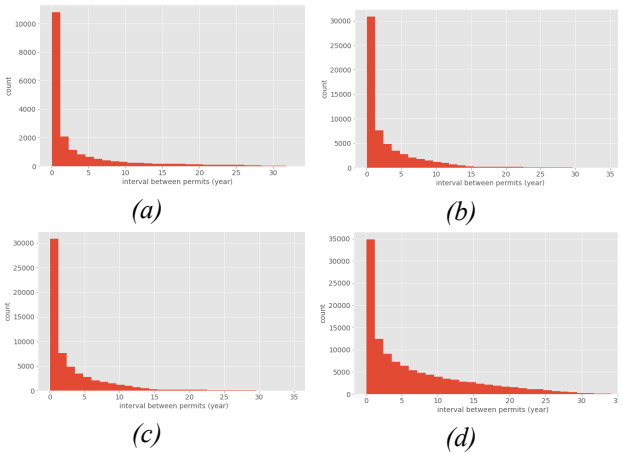


Figure 8 Distribution of time intervals between (a) mechanical retrofit, (b) electrical retrofit, (c) plumbing retrofit and (d) building retrofit for each building

Table 4 Statistics of time intervals between each type of permits for each building

	BUILD-ING	ELEC-TRICAL	MECHAN-ICAL	PLUMB-ING
Mean	3.76	3.22	4.55	6.70
Median	0.91	1.20	1.56	3.96
Max	34.00	35.49	34.96	35.28

It can be observed from the histogram that the majority of the intervals are around 1-2 years. Most of the buildings have a retrofit (changes requiring a permit) every two years. The interval for mechanical retrofits is generally the shortest, the median of which is less than one year, while the building retrofit is not that frequent, which happens around every four years.

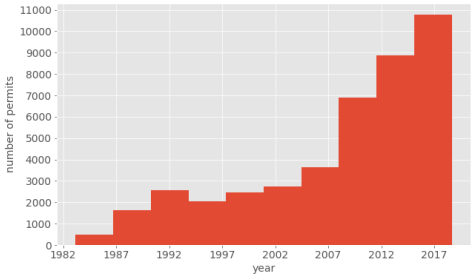


Figure 9 Distribution of the most recent time completing a mechanical work for each building

To infer the potential retrofit in the future, take mechanical retrofit as examples. The last time of the building having a mechanical permit is illustrated in Figure 9. Only around 24% (42,092 and 41,522) buildings have had mechanical work or window work in history. Much of the last time for mechanical retrofit falls in the recent ten years. This implies an opportunity for mechanical / window retrofits in the future.

DISCUSSION

While advanced machine learning methods are used to infer the work type of each permit, the prediction accuracy can be hurt due to several reasons. Firstly, the training data we used is permits from other cities. The extracted language features inevitably have differences from those of SF dataset Besides, the ways of defining the work type vary a lot. For example, some cities have a specific label for HVAC work, but many cities just use a broader type "mechanical" to include all work such as HVAC, equipment, and so on. Secondly, the original training data can be mislabeled, because some permits are even difficult to classify by human due to the ambiguous information. To briefly investigate the quality of the training data, 100 randomly-sampled permits from the training dataset are manually labeled to compare with the original labels. The labeling differences for the electrical, mechanical, plumbing and building permits are 6%, 19%, 24% and 26% respectively. It's worth mentioning that the human labels are from the knowledge of the author, which is not necessarily the ground truth. The differences are produced by the gaps between the interpretation. For example, it's difficult to decide what type of work is included in the permit by the description "tenant finish out", which explains the relatively low accuracy of "Building" permits on the test set to some extent. To improve the results of classifying the work type, it might be helpful to get the domain-specific word-embeddings from the permit datasets. For example, use the unlabeled SF dataset to do language modeling to get the customized word embedding.

Furthermore, knowing the general work type is just the priliminary analysis of the permit contents. However, bigger challenges exist to further understand the detailed work of the permit, such as whether the mechanical work is relocating, upgrading or repairing the system, because there is no such label to do supervisory machine learning, and many permits are too brief to provide any detailed information.

Issues mentioned above can all be benefited from a better schema for permits dataset. Some efforts ("Building & Land Development Specification (BLDS) Data Specification," n.d.) have been made to standardize the building permit schema across different cities, but few cities comply with the standard. However, small efforts in recording the data can bring much more value to the datasets. For example, if the type of work and the detailed work content can all be recorded in a standardized format with categorized attributes, such as "action" (repair, replace, upgrade, remove, etc.), "building_component" (window, HVAC system, sink, etc.), it would be much easier to use the information for building simulation. Besides, if all cities share the same data schema, a pretained machine learning model for analyzing the permits dataset of one city can be better transferred to other cities with minor fine-tuning efforts.

Overall, to fully utilize the permits data, the future work includes first further understanding the detailed actions of each permit, and then combined with the corresponding building codes, mapping the actions to the changes of certain building properties that can be directly used in building simulation model. This will provide a scalable method to more accurately infer the building's efficiency level, compared with the current common practice that infers mainly from the built year of buildings. Other interesting insights with respect to the big picture of the urban development, such as the stimulus of the building codes or energy policy can also be extracted with detailed understanding of the data. The author also expects a shared schema of permits data can be widely-adopted in the future, which will make the method really scalable across the states.

CONCLUSION

This study analyzes the public permits dataset of San Francisco to extract useful information for building retrofit analysis. Especially, we explore the way to use text mining methods on the permit description texts. The state-of-art NLP model, BERT, fine-tuned with similar datasets from other cities, is used to classify the type of work of the San Francisco permits. Most of the completed permits are filed for interior, envelope or structural work in residential buildings. Only around

10% of the permits are for mechanical work. For each building, the average time interval between mechanical retrofits is around 3.8 years. Only 24.2% of buildings have mechanical permits in the history, and more than half of them haven't had any mechanical permits in the recent four years, which implies the opportunity for a mechanical upgrade in the future. Although the analysis is conducted on the SF building permits, the method can be applied to any city with a similar permit dataset.

ACKNOWLEDGMENTS

The work was sponsored by Daikin Industries, Ltd, Japan. The work is also supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Office, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

Building & Land Development Specification (BLDS) Data Specification. (n.d.). Retrieved from https://permitdata.org/about.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. (Mlm). Retrieved from http://arxiv.org/abs/1810.04805

Eirinaki, M., Dhar, S., Mathur, S., Kaley, A., Patel, A., Joshi, A., & Shah, D. (2018). A building permit system for smart cities: A cloud-based framework. *Computers, Environment and Urban Systems*, *70*, 175–188. https://doi.org/10.1016/J.COMPENVURBSYS.2018.03.006

Gunay, H. B., Shen, W., & Yang, C. (2019). Text-mining building maintenance work orders for component fault frequency. *Building Research and Information*, *47*(5), 518–533. https://doi.org/10.1080/09613218.2018.1459004

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). *A Convolutional Neural Network for Modelling Sentences*. 655–665. https://doi.org/10.3115/v1/P14-1062

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Retrieved from http://www.aclweb.org/anthology/D14-1162

Stevenson, J. R., Emrich, C. T., Mitchell, J. T., & Cutter, S. L. (2010). Using Building Permits to Monitor Disaster Recovery: A Spatio-Temporal Case Study of Coastal Mississippi Following Hurricane Katrina. *Cartography and Geographic Information Science*, *37*(1), 57–68. https://doi.org/10.1559/152304010790588052

Vert, S., & Vasiu, R. (2017). Augmented Reality Lenses for Smart City Data: The Case of Building Permits. In Á. Rocha, A. M. Correia, H. Adeli, L. P. Reis, & S. Costanzo (Eds.), *Recent Advances in Information Systems and Technologies* (pp. 521–527). Cham: Springer International Publishing.