Highlights

Current Practices and Infrastructure for Open Data based Research on Occupant-centric Design and Operation of Buildings

Mikkel B. Kjærgaard,Omid Ardakanian,Salvatore Carlucci,Bing Dong,Steven K. Firth,Nan Gao,Gesche Margarethe Huebner,Ardeshir Mahdavi,Mohammad Saiedur Rahaman,Flora D. Salim,Fisayo Caleb Sangogboye,Jens Hjort Schwee,Dawid Wolosiuk,Yimin Zhu

- Research on occupant-centric building design and operation can benefit from open data.
- New methodology and tools are essential to facilitate sharing and use of open data.
- Data anonymization can address privacy/ethical issues concerning open data publication.

Current Practices and Infrastructure for Open Data based Research on Occupant-centric Design and Operation of Buildings

Mikkel B. Kjærgaard[a,*], Omid Ardakanian[b], Salvatore Carlucci[c,d], Bing Dong[e], Steven K. Firth[f], Nan Gao[g], Gesche Margarethe Huebner[h], Ardeshir Mahdavi[i], Mohammad Saiedur Rahaman[g], Flora D. Salim[g], Fisayo Caleb Sangogboye[a], Jens Hjort Schwee[a], Dawid Wolosiuk[i] and Yimin Zhu[j]

[a]*SDU Software Engineering, University of Southern Denmark, Odense, Denmark*
[b]*Department of Computing Science, University of Alberta, Edmonton, Canada*
[c]*Energy, Environment and Water Research Center, The Cyprus Institute, Nicosia, Cyprus*
[d]*Department of Civil and Environmental Engineering, Norwegian University of Science and Technology, Trondheim, Norway*
[e]*Department of Mechanical and Aerospace Engineering, College of Engineering and Computer Science, Syracuse University, Syracuse, USA*
[f]*School of Architecture, Building and Civil Engineering, Loughborough University, Loughborough, UK*
[g]*Computer Science and Information Technology, School of Science, RMIT University, Melbourne, Australia*
[h]*Energy Institute, Bartlett School of Environment, Energy and Resources, University College London, London, UK*
[i]*Department of Building Physics and Building Ecology, Vienna University of Technology, Vienna, Austria*
[j]*Department of Construction Management, College of Engineering, Louisiana State University, Baton Rouge, LA, USA*

ARTICLE INFO

Keywords:
Open Data, Data Publishing, Data Use, Occupant Behavior, FAIR Data, Ontology, Anonymisation, Metadata Schema

ABSTRACT

Many new tools for improving the design and operation of buildings try to realize the potential of big data. In particular, data is an important element for occupant-centric design and operation as occupants' presence and actions are affected by a high degree of uncertainty and, hence, are hard to model in general. For such research, data handling is an important challenge, and following an open science paradigm based on open data can increase efficiency and transparency of scientific work. This article reviews current practices and infrastructure for open data-driven research on occupant-centric design and operation of buildings. In particular, it covers related work on open data in general and for the built environment in particular, presents survey results for existing scientific practices, reviews technical solutions for handling data and metadata, discusses ethics and privacy protection and analyses principles for the sharing of open data. In summary, this study establishes the status quo and presents an outlook on future work for methods and infrastructures to support the open data community within the built environment.

1. Introduction

Recent studies (1; 2; 3; 4; 5; 6), have shown that occupants' presence and actions have a significant impact on energy consumption and thermal comfort in buildings; however, the role of building occupants is not being sufficiently considered to date. The uncertainty of occupants' presence and actions leads to significant differences between the actual and simulated energy consumption (7; 8). Most building energy simulation tools focus more on the physical design factors (e.g., building materials and constructions, technical systems, external weather) rather than interactions between occupants and building's systems and equipment. In addition, several methodologies for building operations and building modeling typically utilizes a fixed operation schedule based on certain rules such as the ASHRAE 90.1 standard, which results in energy waste and occupant discomfort (4). Therefore, more data on human presence and behavioral actions is crucial for efficient management of modern built environments. If data was available this could enable digital twin representations of buildings that could power new data-driven methods for building operation. For example, the knowledge of human presence can be used to provide real-time analytics about space usage, while predictive analytics can leverage information about both occupants' presence and their actions to optimize building operation.

The study of occupant presence covers the two subareas of occupancy detection and occupancy estimation (9; 10; 11). Specifically, occupancy detection concerns binary inference of occupant presence and absence in different parts of an indoor or outdoor space, whereas occupancy estimation concerns determining the number of occupants in that space. Despite this distinction, accurate occupancy detection and estimation are both quite challenging due to occupancy dynamics and variation in the function and type of target sites (e.g., closed rooms, open plan offices, shopping centers, cinema theaters, and public places).

Occupants also do actions (e.g., open doors and switch lights on) and exhibit a wide range of behaviors in different situations. The ability to correlate actions with energy consumption and to identify or forecast particular activity can help minimize exhaustion of unnecessary energy resources.

*Corresponding author address: Campusvej 55, DK-5230 Odense M, T +45 65507965, mbkj@mmmi.sdu.dk
ORCID(s): 0000-0001-5124-744X (M.B. Kjærgaard); 0000-0002-6711-5502 (O. Ardakanian); 0000-0002-4239-3039 (S. Carlucci); 0000-0003-1603-9738 (B. Dong); 0000-0001-5911-2822 (S.K. Firth); 0000-0002-9694-2689 (N. Gao); 0000-0002-1304-4366 (G.M. Huebner); 0000-0003-2320-0112 (M.S. Rahaman); 0000-0002-1237-1664 (F.D. Salim); 0000-0001-9995-758X (F.C. Sangogboye); 0000-0001-9176-2024 (J.H. Schwee); 0000-0002-0015-6000 (D. Wolosiuk); 0000-0001-9176-202x (Y. Zhu)

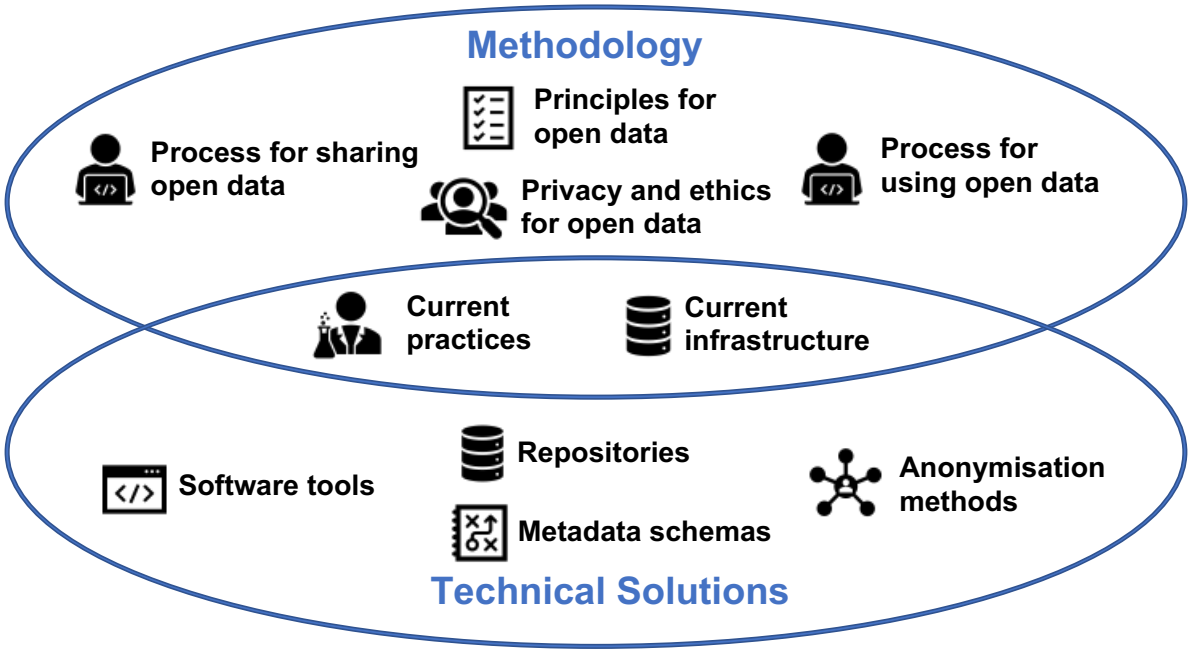

Figure 1: Overview of methodology and technical solutions for open data

This includes all activities that occupants engage in indoors because each activity can influence the energy consumption state.

The concept of open data is still new with relatively sparse definition capturing their essence and purpose. (12) and (13) identified that an open data is characterized by freely available data with limited restrictions with respect to the reuse, republishing, and redistribution of data. (14) defines open data as "non-privacy-restricted and non-confidential data which is produced with public money and is made available without any restrictions on its usage or distribution". Recently, the concept of open data is increasingly expanding from it numerous and concerted outsets mainly from governmental initiatives (15; 16) and it is now receiving increasing attention in many fields in the scientific community. (15) presents a genealogical perspective to the advances in open data. This work provides a reflection of how open data has been utilized as a tool for shaping various governmental and scientific discourse and for ensuring transparency and openness in empirical studies. More specifically, (17) and (18) highlights the importance of applying open data for enabling data-driven models in the cyber-physical space especially in building performance research and applications in relations to occupant behaviors and activities.

Open data on occupant presence and behavior could help researchers improve the understanding of interactions between the occupants and the buildings in different contexts. This can be significant for optimizing building energy use and for providing improved ambiance. One important aspect when collecting and using data on occupant presence and behavior is the possible privacy implications and need for privacy protection (17). Compared to many other types of building data this is a particular important aspect for data on the presence and behavior of occupants.

Given that open data from the public domain could play an important role in building performance research and applications vis-à-vis occupant behavioral research, a better understanding of the benefits and challenges for applying open data could be beneficial for researchers and scholars that are applying them in these endeavors. To enable the research communities to apply and utilize open data, it is necessary to reach a consensus about accepted methodologies and technical solutions. Figure 1 highlights the methodology and technical solutions for the successful application of open data. From a methodology perspective, the community has to establish processes for sharing and using open data, principles for open data, and guidelines for the specific privacy and ethical questions that are raised about open data. From a technical perspective, the community has to establish repositories for sharing data, algorithms for anonymizing data, metadata schemas for assigning semantics to shared data and software tools that track data transformation and make it easy to share and use open data. A key challenge for technical solutions is interoperability: the ability of the individual systems to communicate and exchange information in a meaningful manner. In the intersection between methodology and technical solutions, we have the current practices and infrastructures.

This article reviews current practices and infrastructure for open data-driven research on occupant-centric design and operation of buildings. The article covers related work on open data, presents survey results for existing scientific practices, analyzes the availability and use of open data and associated infrastructure, and investigates handling of data privacy and ethics. In summary, this study establishes the status quo and presents an outlook on the future work for meth-

ods and infrastructure to support the open data community within the built environment.

2. Background

This section covers the background of open data and current efforts in the built environment. As open data has not yet been extensively applied to occupant presence and behavior data the background will mainly focus on other types of data.

Gray et al. (15) traced the genealogy of open data and found several threads of evolution. Mainly, according to Gray's view, the idea of open data arose from debates about data collected by government or public entities and the role of such data in economic growth, innovation, promoting transparency and efficiency of governments. Janssen et al. (14) analyzed the benefits and barriers related to open data and open government. They argued that open data itself had limited value; it was the use of open data that created value (14). Therefore, open data needs an infrastructure to support its use, such as management, discovery, curation, analysis and visualization. Because publishing data requires resources to sustain, it is simply more than just putting data online. Benefits of open data include political and social benefits (such as transparency and equal access to data), economic benefits (e.g., stimulation of innovation and creation of new economic sectors), and operational and technical benefits (e.g., reuse of data). Barriers include institutional factors (e.g., regulations and legislation, security, external safety (19)), task complexity (e.g., inability to convert data, improve storage system), use and participation (e.g., ownership, liability, privacy), and legislation (14).

In an attempt to answer a question related to government's position to open data, Rudmark et al. (20) resorted to the idea of digital ecosystems, as "a distributed adaptive open socio-technical system with properties of self-organization, scalability, and sustainability" (21). For example, through the study of the public transport industry in Sweden, Jansen et al. (21) found that open data created significant additional value due to its reuse by existing digital ecosystems and the creation of potentially new ecosystems, where government as a data provider took a peripheral position, as opposed to assuming an active role in managing open data. Zhu et al. (22) reported that after the US Geological Survey (USGS) made the Landsat data publicly available on the Internet in 2008, there was a significant increase in the use of Landsat data in public and private domains. Reported values include the reuse of data, economic benefits, international collaboration and commercial cloud computing services.

2.1. Benefits and challenges around open data

Reproducibility is key in science and some advocates for open data argue that too few research studies are reproducible. Open data offers researchers a solution to the problem of reproducibility and an opportunity to expand their observations, which may accelerate new discoveries (23). Open data in science requires advocacy and coordination (12). Some scientific disciplines such as ecology (24) are interdisciplinary by nature and have benefited from open access to data. Measures must be developed such as policies and guidelines to guide researchers who are not familiar with the open data process (23). For reproducibility to be achieved the open data should be interoperable, i.e. it should be understandable and usable for researchers who did not undertake the original data collection (see Section 8).

Reichman et al. (24) emphasized the need for standardizing metadata development, the reproducibility of analysis and rewards for sharing data, in addition to well-curated integrated open data. This work describes a technical process of DataONE which enables federated access to ecological data. This process includes the data acquisition workflow (data acquisition, quality assurance, metadata and semantics, and data deposition), data federation, discovery and access, and data analysis workflow (integrate and transform, analysis, modeling and visualization). Reichman et al. (24) further highlighted the data challenges in ecological informatics and these involves dispersion, heterogeneity, and provenance. Dispersion highlights a data collection process involving many individuals with different experience/data collection protocols and across a large number of geographic locations. (24) highlighted that dispersion makes it difficult to researchers to discover data. Even when data are discovered, researchers often face the challenge of heterogeneity because data are collected for different purposes and with different protocols. Finally, data typically go through steps of transformation before meaningful results are observed. Such transformation process needs proper documentation for reproducability. Lastly, Reichman et al. (24) highlighted a number of social and cultural barriers (e.g., the need for a rewarding system).

Chen et al. (25) argued that the open data practice is not enough to ensure the reproducibility of scientific results. "It is also essential to capture and structure information about the research data analysis workflows and processes to ensure the usability and longevity of results". To achieve this objective, "research communities may start by using open data policies and initiating dialogues on data sharing, while embracing the reproducibility and reuse principles early on in the daily research processes." Open data allows reusing data for research, education, and training and it offers researchers a solution to the challenge of reproducibility and an opportunity to expand their observations, which may accelerate discoveries (23). For example, datasets could be combined into a larger one, allowing greater statistical power. The main personal benefit lies in the increased visibility of one's research; most data repositories issue a digital object identifier which means the dataset can be cited easily and, most likely, previous papers would also be cited. Also, nowadays journals exist whose main article type is a published data set, for example, Nature Scientific Data's data descriptor.

These benefits outweigh the perceived disadvantages of open data. There can be concerns about the privacy implications for the subjects, when sharing open data. Typically addressed by using pseudoanonymizing and anonymization on the data prior to publication. Furthermore, most relevant privacy laws and regulations be identified and considered; e.g.

the releasing part needs to consider relevant local privacy laws, e.g., if monitoring EU citizens consider the General Data Protection Regulation (GDPR) (26), if monitoring California citizens consider the California Consumer Privacy Act (CCPA) (27), or in Australia, consider the Australian Privacy Principles (APPs) (28). Most likely data would need to be deidentified anyway to be shared within a team or even comes deidentified already (such as online surveys). Some studies result in very large data sets. While many repositories are free to use, there is sometimes an excess charge for very large data sets. However, in the light of what research actually costs, the deposition cost is often negligible. The drive to publish in competitive academic environments might also play a negative role: data collection can be a long and expensive process and researchers might fear premature data sharing may deprive them from the rewards of their efforts including scientific prestige and publication opportunities. However, it is possible to mitigate by delaying data sharing until having done all planned research.

2.2. Open Data and the Built Environment

Only recently open data is being applied to support occupant-centric building design and operation. For example, no subject specific data repositories exist so far. However, the creation of data sets and the availability of open data about the built environment are emerging boosted by the concept of smart cities. Dixon et al. (29) argue that the smart city concept is fostering the access to built-environment open data hubs that are getting available in an increasing number of cities. Typically these open data hubs include data about transport, energy, land use and property, but often no data about the actual behavior of people is made available, commonly for privacy issues. Several examples are available of open data sets that have been established recently to support occupant-centric research but most of them without explicit data on occupant presence and behavior. For example, Miller et al. (30) discussed the Building Data Genome project and an open non-residential data set of building characteristic and electrical meter data. The data set combines existing open source data and additional data gathered by the authors but do not explicitly include occupant-centric data. Roth et al. (31) discussed the use of open data for developing a new urban building energy model. Their open dataset includes the 2016 energy data of the New York City as local Law 84 makes public access possible. The building data were collected from the Primary Land Use Tax Lot Output dataset that is publicly available. The authors used the Archetype hourly building loads of the DOE reference building simulation models. Barker et al. (32) described the development of two open data sets for residential homes that include data about occupant presence based on motion sensors. These efforts range from curating and publishing data to developing novel applications using open data.

The motivation of pushing governments for open data about the built environment is multifold, such as (i) increase transparency and accountability of government operations, (ii) support economic development, (iii) foster research and innovation, and (iv) it is a right of taxpayers. Regardless, the open data movement requires strong incentives that can mobilize a critical mass who owns data and values open data. In science, open data has been argued from the perspective of reproducibility and acceleration of discoveries, cost benefits, and sometimes the nature of scientific work. Again, policies and guidelines are essential to create a critical mass within the built environment community. Once the human factors are addressed, the technical and policy factors can facilitate the development of open data communities for the built environment.

3. Open Data Publication and Use

The collection, publication and use of open data include a number of steps. Figure 2 provides an overview from data collection to end use to provide a common frame for the individual sections in the paper.

Data collection (1): Relevant occupant-centric data can be generated by sensors, control points, system logs, or manual observations, as covered by (33). To collect occupant-centric data, many different types of sensors and data sources are available. The broad sensor types include image-based, threshold and mechanical, motion sensing, radio-based, human-in-the-loop, and consumption sensing. The data might be stored on local sensor storage and then later copied to a repository or transferred to a repository directly via wireless sensors, gateways, building automation systems, or internet-enabled sensors. Data points have to be associated with metadata that explains their origin, for example, based on information from databases or text manuals. We refer the reader to (33) for more details.

Data Cleaning and Normalization (2): Prior to transferring data into a structured repository it must be examined in view of consistency, completeness, and plausibility. Data cleaning and normalization includes processing data to treat omissions and anomalies and normalize or convert raw data as needed. Section 5.1 will cover possible formats and implementation frameworks for data processing.

Local Repository and Metadata (3): Data is transferred to a repository of the collecting organization and made available from here, including metadata (covered in section 6.1). The internal repositories can be constructed in different ways depending on the size and access pattern for data. In a simple setup, the repository might consist of directories of plain files and in an advanced setup, might be a number of databases, e.g., time-series or graph databases.

Data Anonymization and Publication (4): Data is published, and the necessary processing is applied, including anonymization. Methods for anonymization to address ethics and privacy protection are covered in section 7.2.

Open Data Repositories (5): Data is made available in an open data repository that can provide data by relevant means

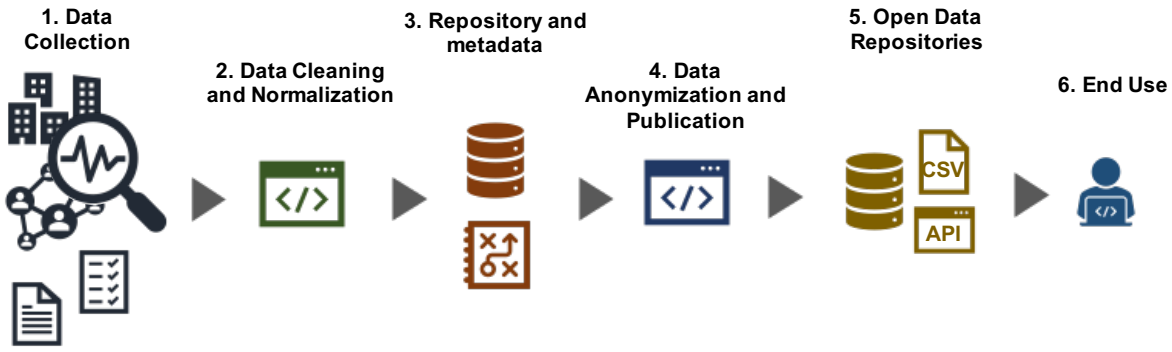

Figure 2: Overview of process for collection, publication and use of open data

(covered in section 5.2), for example, as downloadable CSV files or dynamic data access via an API. The data can both be unstructured data for human consumption with limited additional information about the data or structured data created for computer consumption that can be queried to extract relevant subsets.

End Use (6): The end user can then finally process the data by relevant software tools to do the research of their focus, for example, software analytics combining, and correlating data about occupants, building structure and subsystems to understand how occupants interact with buildings and how buildings respond to occupants' behavior and actions.

4. State-of-the-Practice for Open Data

The review of the benefits and challenges of open data and use by other sciences have highlighted that some fields have well developed ecosystems and practices for open data. This paper focuses on the interdisciplinary field of occupant-centric building design and operation. To capture the current state-of-the-practice for open data within this field we have designed a questionnaire for researchers in the field and administered it throughout the web to representative communities.

4.1. Method

The IEA EBC Annex 79 is a recent effort into research on occupant-centric design and operation of buildings established by IEA. Invitations to participate in the questionnaire was sent out in march 2019 to the 116 participants on the mailing list of the Annex 79. Participants were also motivated to share the survey with other researchers in their research groups and relevant mailing lists. Responses were collected until August 2019.

The questionnaire contained questions in these areas:

1. Area of expertise,
2. Use of data formats and software tools,
3. Use of open data,
4. Barriers for using open data,
5. Sharing of open data,
6. Barriers for sharing open data.

In total nine questions were asked where five were answered with options on a choice list and four were answered with an open textbox. The options on the choice lists where based on the authors knowledge on open data and barriers mentioned by previous work including (34). All choice lists also included an option "other" that allowed respondents to add additional choices to the lists.

4.2. Results

One hundred and eight respondents opened the questionnaire of which 34 respondents completed the survey. Because no complete list of all active researchers in the field exists, an estimate for the complete population is not available, and given the low number of possible respondents the study will not be able to have a proper number of responses for statistical analysis. Therefore, the study is limited to provide only indications on current practices. Figure 3 reports results for area of expertise, data formats and software tools. The results for area of expertise shows that 70% of the respondents belong to the area of engineering followed by 16% information technology and 12% architecture. In terms of the use of data formats almost all respondents were comfortable using comma-separated files and 56% where comfortable of using database technology. For structured data formats 42% and 33% were comfortable using JavaScript Object Notation (JSON) and eXtensible Markup Language (XML), respectively. Only 7% were comfortable with Resource Description Framework (RDF). For software tools 84% were comfortable using spreadsheets, 67% mathematical programming tools and 65% general programming languages. Only 28% and 19% were comfortable using specialized data visualization and data plotting tools, respectively.

For the text questions on the use of open data 12 respondents reported that they have knowledge of open datasets. All of these also reported to have used open datasets in their research. Seven respondents have answered that they have also shared open datasets. The participants with positive answers included references to datasets and papers in their answers. This is actually a high percentage of the respondents. Taking the journal "Building and Environment" as a comparative example only 13 out of the 1310 papers published in 2018 and 2019 until end of October mentioned the phrase

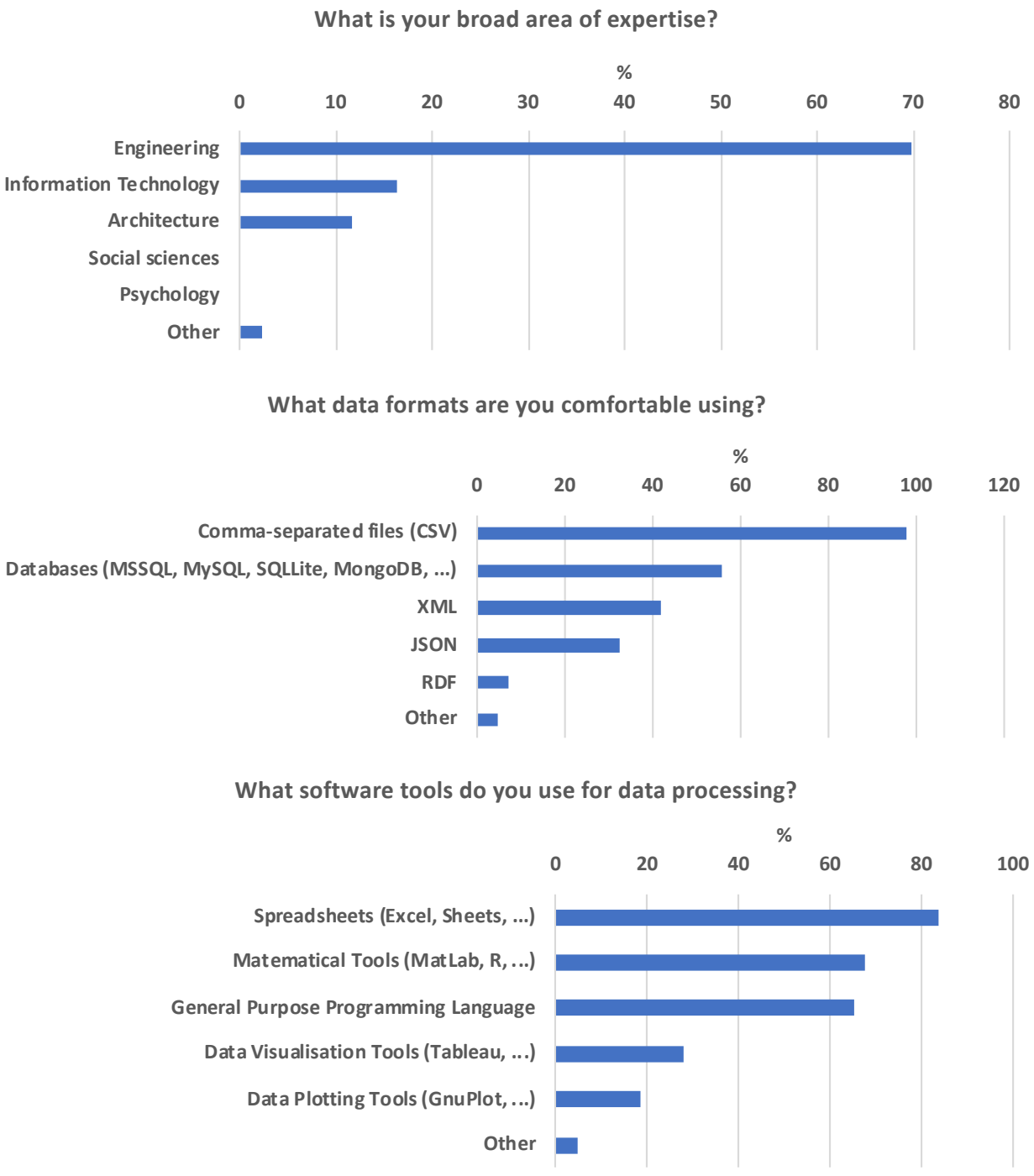

Figure 3: Survey questions and results

"Open data". Currently, the website of the journal also provides links to 13 datasets deposited at Mendeley Data Repository. The provided information by respondents are covered in Section 8 which analyze existing open datasets and their use.

Figure 4 presents the responses for questions on barriers for using and sharing open data. 78% of the respondents report knowledge of available datasets as the largest barriers followed by lack of documentation (53%). Other barriers with many responses include high overhead of using data (49%) and lack of experience (25%). In the "other" category for additional barriers the respondents' mentioned that released datasets might not have a high enough sampling rate of sensors for specific studies. For sharing data the main barriers reported by respondents are ethical and security concerns (56%), time consumption for preparing data for release

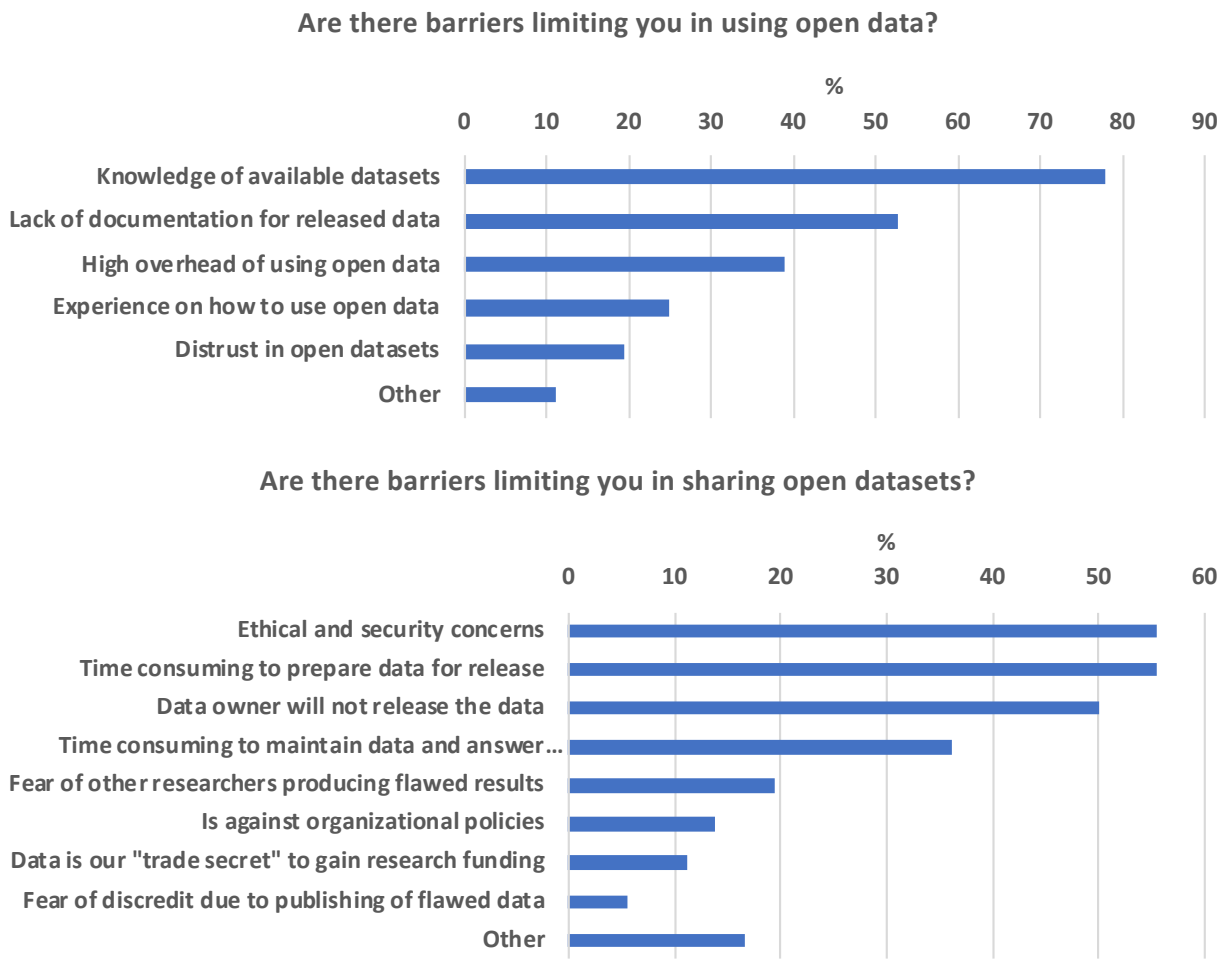

Figure 4: Survey questions and results

(56%) and maintenance (36%), and data owners who do not like to share data (50%). Under the "other" category respondents mentioned lack of incentives as a barrier.

The study results for current practices highlight that many of the respondents know programming, mathematical tools and different file formats. Therefore, many researchers have the skills needed to do research based on open data. However, the results also highlight several barriers for using open data including knowledge of datasets and lack of documentation for datasets and for sharing data including time consumption and concerns limiting release of data. Therefore, efforts are needed to address such barriers to develop a successful ecosystem for open data within the area of occupant-centric building design and operation.

5. Infrastructure for Open Data

Sharing and using data require software tools to acquire, process, store and transfer data. In this section we cover the existing tools and platforms, and list the data repositories that supports the publication of open data.

5.1. Software Tools

A plethora of open-source software toolkits and platforms have been developed in recent years to facilitate storage, processing, and analysis of a large volume of time-series data and metadata. Most of the existing software platforms are general-purpose and are adopted by many in academia and industry. Examples of these platforms are OpenRefine, Pandas, scikit-learn, Keras, and Cloud AutoML.

Specialized software platforms have also been developed in recent years to transform and process energy-related datasets collected from residential and commercial buildings. For example, NILMTK (35) is a toolkit designed for evaluating and benchmarking energy disaggregation algorithms. It imports and transforms open datasets suitable for energy disaggregation. However, little work has been done to date to develop software platforms that are suitable for aggregating, structuring and analyzing occupancy-related data collected from the built environment.

Mortar (36) provides an application execution environment and an API which can be used to develop and evaluate building applications across a large number of buildings represented in the Mortar dataset. The platform performs

basic data cleaning operations (hole filling, filtering, and aggregation). Thereby, Mortar represents an open testbed for portable building analytics containing time series data generated by sensors and control points located in over 100 buildings, and metadata for these buildings.

ODToolkit (37) is an open-source building-occupancy data processing toolkit written in Python (available on GitHub). It is comprised of five main modules for data ingestion and aggregation, preprocessing, analysis, evaluation and plotting. The first module can pull in a subset of the existing open datasets from various repositories, and converts them to a unified data format (i.e., NumPy multidimensional arrays) so that they can be efficiently stored and retrieved. The preprocessing module tackles various data quality issues in a number of steps. This involves detecting and removing outliers, imputing missing data points, unifying the sampling frequency of different features in different datasets, and semi-automated mapping of feature names in different datasets to standard names. The data analysis module allows for training and testing a large suite of supervised learning models and semi-supervised domain-adaptive models provided in this toolkit or added by users. The evaluation module includes a suite of metrics for comparing the models, and the plotting module provides various methods to create plots and showcase the results. ODToolkit fully or partly addresses some of the issues that are not typically addressed by general-purpose data processing and analytics platforms (e.g., ontology mapping). It has a modular design and can be extended by users to incorporate new data sets, algorithms, and metrics.

obFMU (38) provides a simulation module for occupant behavior. The simulation module integrates with EnergyPlus and other building performance simulation tools. The parameters of the simulation are specified in data specification files in the obXML format (discussed in Section 6.1). Thereby, the simulation module enables the simulation of occupant behavior based on data that could come from open data sources specifying the behavior of different building environments.

PAD (17; 18) is an example of a privacy preserving data publishing software framework that presents another approach for enabling the sharing and use of open data. This approach involves the re-use of standard privacy preserving methods such as *k*-anonymity for anonymization and for publishing privacy sensitive dataset obtained in cyber-physical spaces. PAD further expands on this idea to provide a customization function that can extract specific interests of given data recipient or data-driven application and incorporate that interest in the data anonymization and publication process to improve the utility of the processed open data.

While these software tools are good examples of developments that support the publication and use of open data, they seems to present a more solitary and specialized approach for publishing or utilizing open data. This consequently will require intending data analyst or recipient to ingrain themselves in the domain specific interfaces presented by each software tool to utilize and extend the functionalities for their domain applications.

5.2. Repositories

The registry for research data repositories[1] provides a list of over 2,000 repositories covering different academic disciplines. While some the repositories are openly accessible, access to a number of institutional webpages are restricted. Fairsharing[2] also provides a comprehensive list of databases, with information on subjects, domains, taxonomies used, etc. Nature Scientific Data publishes a list of recommended data repositories on their webpage[3]. Whilst there are no recommended data repositories for occupant behavior in buildings, repositories for social sciences, physics, and general use are listed. Figshare[4] and the Open Data Framework[5] are two commonly used general repositories. Both – as many other – allow generating a digital object identifier (DOI), meaning the data can be cited using traditional citation methods. Many domain specific repositories are also available where OpenEI[6] is a data repository for various energy usage related database in the U.S. and internationally. The repository includes open data on geothermal, wind farms, utility rate and occupant behavior data in buildings. The latter only includes a few datasets from field measurements, such as a one-year dataset collected from a single office, a one-month dataset from four residential houses, and synthetic generated occupant behavior data from agent-based models. It provides API access so that users can easily get the data from the database, but it does not generate a DOI. In addition, there are portals or platforms for managing open data such as CKAN[7] that are available as open source software so an organization can setup their own repository for open data.

6. Handling Metadata

The background literature and survey results highlight the need to overcome the barriers due to lack of documentation and standardization. A solution to these problems is to digitally define the semantics of data also known as "metadata". The metadata describes the context, content and structure of data.

Defining and improving the metadata of a dataset greatly improves its interoperability. A well-formed interoperable dataset will have both concepts and data described in a logical structure with clear descriptions and relationships defined for all terms involved. This enables both humans and computers to be able to read and understand the dataset without any prior knowledge of its format or structure.

6.1. Metadata Schemas

Metadata might be structured by a schema, which describes the structure of data or by ontologies, which in ad-

[1] http://www.re3data.org/
[2] https://fairsharing.org/
[3] http://www.nature.com/sdata/policies/repositories#general
[4] https://figshare.com/
[5] https://osf.io/
[6] https://openei.org/wiki/Main_Page
[7] https://ckan.org/

dition to structure also can define objects and semantic relationships. In the following, a number of schemas and ontologies are described: 1) gbXML focusing on building construction; 2) IFC focusing on building construction; 3) BPD focusing on building performance data; 4) Brick focusing on building data at large and 5) DNAS/obXML focusing on occupant behavior simulation. These five instances of schemas/ontologies offer – to various extents – features that may be of relevance toward the representation of occupancy-centric information. Other examples of building-related ontologies/schemas include BuildingSync[8], Project Haystack[9] and the Building Topology Ontology[10]. BuildingSync is a schema that focuses on buildings' energy audit data. Project Haystack introduces data schema for building systems and equipment. The Building Topology Ontology allows for representation of relationships between building's zones and components.

gbXML[11]: Green Building XML is developed by Green Building Studio with the support of the California Energy Commission Public Interest Energy Research (PIER) Program, and the California Utilities. gbXML currently facilitates the exchange of data among CAD tools such as Autodesk Revit and energy analysis software such as eQuest. It can represent different objects to describe a whole building including geometry, HVAC system, and schedules. The latest schema is version 6.0.1, 2017. The current occupant behavior is represented as "PeopleNumber" and "PeopleHeatGain" inside "Space" element. It has attributes "peopleScheduleIdRef" to link with "Schedule" element that defines the actual schedule. gbXML has been used to mapping material properties to facilitate lighting simulation, thermal properties for building energy modeling, early design decision support, and building retrofit analysis (e.g. as prestended by (39)). However, gbXML has not been used to represent behavior aspects of the occupancy such as window opening, light switch or thermostat setpoint changes due to the lack of necessary XML elements in the current version.

IFC: Industry Foundation Class (IFC)[12] aims to provide a universal metadata basis for process improvement and information sharing in the construction and facilities management industries including smart buildings. IFC is certified by ISO in 2013 as an international standard-ISO 16739-1:2018. The data schema of IFC is defined in EXPRESS and can be generated as an XML file as well. The current version of IFC is 4.2. IFC has been widely used to represent metadata for construction industry to exchange various information from CAD design to scheduling and cost estimation. A few studies using IFC to exchange geometry information between CAD tools and building energy simulation models (e.g. (40)), and lighting simulation models. Most recently, IFC has been applied to the building fault detection and diagnostics by (41). The occupant presence potentially could be represented by IFC as the "IfcTimeSeriesValue" and attached to the "IfcOccupant". However, like gbXML, the behavior of occupants cannot be represented.

BPD Ontology: Building Performance Data Ontology was introduced in a number of publications (see, for instance (42)) to address the need for a general, robust, and versatile data structure for building monitoring data, including data categories pertaining to indoor and outdoor environments, control systems and devices, equipment, energy, and occupancy. It was subsequently extended to cover general building performance data including building performance indicators ((43)). The BPD Ontology captures the multi-faceted nature of building performance data in terms of a general schema. Thereby, salient characteristics of performance data – within hierarchically ordered sets of categories and subcategories – are documented via specification of variables attributes (label, value, type, unit, temporal, spatial, and frequency features, as well as data source and auxiliary information). The BPD ontology facilitates, among other things, the organization and representation of occupancy-related data, including time series of monitored and/or simulated values of variables that capture occupants' presence and behavior in buildings.

Brick: The metadata model named Brick (44) was presented with the stated goal of expressing all relevant relationships in a building. Brick cover buildings and their components including sensors (e.g., temperature or light level sensors), subsystems (e.g., ventilation and heating) and their relationships. Brick represents relevant building information as a graph expressed using a Resource Description Framework (RDF) triplestore which supports definition of classes and subclasses of entities (i.e., the nodes and edges of the graph in the form subject-predicate-object, e.g. "room12 is-on floor-3".). RDF supports namespaces for organizing sets of triples. A Brick model can be queried using the SPARQL language. A SPARQL query defines a pattern based on relationships between entities and names of key entities to extract. For each match of the pattern found in the model the result of the query will contain the concrete values for each named entity.

DNAS/obXML: The DNAS (Drivers, Needs, Actions and Systems) ontology was developed to address the need for a consistent representation of energy-related occupants' behavior in buildings; particularly with regard to such behavior's potential influences on buildings' energy use (45). It is based on four human-building environment interaction framework components (46). These include the drivers of behavior (external environmental factors), the needs of the occupants (physical and non-physical comfort requirements), the actions carried out by occupants (interactions with systems to satisfy needs), and the building systems acted upon (equipment, mechanisms to interact with). The DNAS ontology was implemented in an XML schema called obXML (47). The occupant behavior is represented by three main elements. The *Buildings* element puts the occupant in a spatial context. The *Occupants* element captures detailed information about an occupant. The *Behaviors* element parents

[8] https://buildingsync.net/
[9] https://project-haystack.org/
[10] https://w3id.org/bot#
[11] https://gbxml.org
[12] https://www.iso.org/standard/70303.html

the aforementioned DNAS framework components. The element tree of each of the *Behaviors* child elements attempt to capture the stochastic nature of occupants' behavior.

While each scheme has its own strength and weakness, there are few studies comparing or integrating them within the same context/application. Dong et al. 2007 (39) first conducted a detailed investigation and comparative study of the differences between IFC and gbXML in terms of their data representations, data structures and applications. The study selected gbXML due to its flexibility and developed a seamless data integration platform between a CAD model (i.e., REVIT) and lighting simulation software (i.e., Radiance) to support concurrent design of high performance buildings. In addition, in order to support building design, obXML is potentially can be integrated with gbXML, however, the work is still in progress (48).

6.2. Populating metadata for a dataset

After selecting a metadata format the next step is to populate the metadata for a building or dataset. Depending on the setting more or less digital information might be available for the data publisher. Therefore, at the one extreme physical inspections of equipment and manuals might be necessary to establish metadata. At the other extreme the data might already be available in digital form and only need to be processed or translated to the correct format. For example (49) studies the mapping from IFC to brick metadata. However, if metadata is not available in digital form it can be costly to reconstruct it. Therefore, research has explored methods to providing metadata by automatically tagging data based on learned similarities to other data by machine learning. There are mainly two categories of methods in the data labeling landscape: using existing labels of already labeled data or using crowd-based methods. Using existing labels belong to the semi-supervised learning where the idea is to exploit the existing labeled data to predict the likely label of new data. Automated machine learning (AutoML) is an artificial intelligence-based solution for data tagging (50). With AutoML techniques, the labels of data can be learned automatically by self-tuning and autoconfiguration of machine learning models. Another approach is crowd-based method, e.g., active learning which aims for carefully selecting the right examples to reduce the cost for labeling, crowdsourcing techniques where there can be many workers (non-experts) in labeling. Crowd-based methods focus on investigating the task assignment for workers, selecting the interface, and ensuring high quality labels. (51) proposed a collaborative crowdsourcing technique for labeling, which employs crowds to identify uncertain types of data and create rich structures for post-hoc label decisions. (52) proposed to semantically label a physical space with categorical information from DBpedia in order to learn the contextual similarity between the queries and physical space. (53) proposed automatic semantic labeling using machine learning techniques. They mapped attributes to the DBpedia and used similarity metrics as features to compare against labeled domain data. Then a matching function could be learned for inferring the correct semantic labels for the data. Plaster was proposed by (54) as a framework for implementing metadata normalization methods. Recently, transfer learning has received more and more attention because it is considered to be beneficial in saving the cost of tagging, and has shown the possibility of improving the tagging performance when tagging data is sparse. Transfer learning aims to improve the process of learning on a target problem by using the knowledge gained from the training examples in a source problem related to the target one. Transfer learning for tagging data has been applied in various fields such as cross-domain collaborative filtering by (55). However, more work is required to apply this for metadata for occupant-centric building data.

7. Privacy and Ethics for Open Data

In occupant presence and behavior studies, collected data may contain personal information (e.g., gender, age and behavior patterns) from participants and should be considered as sensitive data. This conflicts with the intention of sharing the data as open data. Typically, there are some limitations to the reuse of sensitive data (56). Firstly, researchers are expected to obtain informed consent from participants for the use of collected data. The consent documentation should contain (1) the level of consent for the future use of data, and (2) explicit information on the data to be held in a form that is identifiable, non-identifiable, or re-identifiable. Secondly, researchers should make sure to protect the participant's privacy by de-identifying data when needed.

7.1. Ethical Considerations

In occupant presence and behavioral studies, ethical conduct is important to preserve the individual's privacy and avoid any potential harm from participation in research, especially when considering the potentially high level of personal interaction in the indoor environment. When data is shared as open data, this adds an additional dimension to these considerations. Ethical considerations are similar in different countries though the management process may be country-specific (57). In general, the ethics committee will review all research involving human participants to ensure it is ethically acceptable with minimal risk to participants. The level of review will depend on the degree of risk involved in the research (e.g., whether the participants are identifiable, less than 18 years old, belong to a minority group). Before the occupant-centric studies, ethics protocols and informed consent must be approved. For each research proposal related to occupant presence and behaviors, the researcher should demonstrate that the research has merit and reflects the ethical values of justice, beneficence, and respect for humans (58). The projects related to occupant presence and behavioral analysis are usually complex or long-running. The data collection site may consist of many other people who are not participants. Therefore, the occupant-centric studies should take the necessary steps to protect their privacy as well.

There are some risks specific to occupant behavior research that makes the ethics and data collection more challenging (33). Eventually, these could be an obvious obstacle for any occupant-centric open data release. These challenges could be related to:

- Nature of environmental spaces: The participants in a laboratory experiment consisting of multiple rooms could be easily identified if the characteristics of a particular room (e.g., orientation, shape, level of the floor) is known.
- Within-subject difference: In occupant-centric studies, participants may be asked to provide specific data such as age, gender, height, and weight, which are the required variables for constructing certain thermal comfort models. There is a risk for identifying the participants and causing a data leakage.
- Sensor data collection: The physiological data and movement data can pose privacy risks in occupant studies which allows identifying the specified participant (33; 59).
- Video data collection: Surveillance video data shows potential significant ethical risks (33). For instance, the employee may feel uncomfortable and less productive for fear of their undesirable behaviors are recorded (e.g., lack of presence). .
- Secondary data: Many researchers tend to use the secondary data (i.e., data collected from other studies or sources) which may expose ethical issues related to information leaking.
- Fairness: With the release of new data sets, the data quality may vary tremendously. When the spectacle is prioritized over careful considerations during data analysis, it may result in serious issues such as cultural biases and unsound logic (60).

One solution to mitigate the above risks is to pseudonymise or anonymize the data and delete any information related to the participants (e.g. name, age, height). In case of video/sensor data, researchers must strictly control the access of video/sensor data and may analyze the data at a group level (anonymous individual data is analyzed only when necessary). Regarding the use of secondary data, the amended consent from participants may not be easy to collect. However, the usage of secondary data should go through an approval process from the respective ethics committee to protect the data privacy.

7.2. Anonymisation Methods

There can be two types of identifiers in occupant data that has to be anonymized: direct identifiers such as names, images, or social security numbers, and indirect identifiers, which in conjunction could identify a person, e.g., information on workplace, training, salary, and years of employment. Direct identifiers are usually deleted or at least reduced in precision, e.g., only giving the first values of a postcode. For indirect identifiers, one option is to restrict their range so that outliers that could identify a person are hidden (e.g., turn continuous variables into categorical ones so someone with an atypically high salary would fall into the "high salary category"). For data, that are too detailed, sensitive, or confidential to be made publicly available, secure access environments can be created that restrict access to certain users (e.g., academic researchers) and prevent downloading of data but instead need to be analyzed in a safe online space (e.g., see UK Data Archive Secure Lab[13]). Specific methods for de-identifying exist, which can be applied by first performing a privacy risk analysis and then using a privacy protection method like suppression or a privacy model, such as k-anonymity (61), l-diversity (62), or differential privacy (63) which protected the data agents record-linkage, attribute linkage and probabilistic attacks, respectively. Lastly, all data intended for reuse should have a license.

There are several frameworks developed for protecting time-series data. The PAD (17; 18) framework can protect building-related time series data with the privacy model of k-anonymity. The unique property of the framework is that the data publisher can specify how the data is to be used. These specifications are then considered as part of the anonymization phase of the data. Pythia (64) is an algorithm selection framework which finds the most suitable version of differential privacy, for a given dataset, within the available implementation of the model. When using Pythia, the user is to input the sensitive database, the workload of queries, and the ϵ value for the privacy algorithm. The method first uses feature extraction, upon the inputs, trains the models which are available. The algorithm with the least error, in terms of regret, is selected for the final privacy protection. (65) designed and implemented a replacement AutoEncoder architecture; this can be used for privacy protection of a time-series database. The unique feature of this architecture is that each part of the time-series is put into a disjoint set, which can be sensitive, non-sensitive, and the required utility set. The AutoEncoder is then using feature learning for transforming the sensitive set into non-sensitive data steams.

Selecting a privacy protection method is not a trivial task, and no one method works for every data release. The releasing part needs to consider which part of the to-be-published data is sensitive, if any external data can be used by an adversary for data linkage attacks, and who is monitored in the data. One of the challenges in occupancy centric data is that often the data is time-series, which has a lot of repeating patterns, e.g., an occupant in a privacy office will likely get to the office with in a short time differences each day. Furthermore, an adversary might physically enter a smart building and collect observations about the use, which can be used for breaking the privacy protection. Recent research (66; 67) has identified that state-of-the-practice methods for anonymization can be insufficient to protect the released data against de-identification. One of the findings from (67) is that time-series data aggregated into daily profiles and anonymized with k-anonymity can result in the re-

[13] https://www.ukdataservice.ac.uk/use-data/secure-lab.aspx

leased data not being sufficiently protected against record-linkage attacks. In general, the releasing part needs to apply a privacy protection method that protects against the identified privacy risks.

8. Sharing Principles for Open Data

This section reviews a selection of Open Data datasets in the field of occupant behavior and occupant-centric design for buildings. It establishes the current practice of those researchers and practitioners who are collecting primary data through monitoring campaigns and then releasing the collected data in publicly available, reusable formats. In the field of occupant behavior there is no set format or agreed strategy for collecting and preparing Open Data datasets, so there is considerable variation across the approaches taken. Those who do create and share Open Data datasets can be thought of as pioneers as it often involves significant extra effort to prepare such datasets for publication and this is often done voluntarily without a specific requirement to do so from funders. The purpose of this section is not to criticize or find fault with these efforts, but rather to survey the current practice in publishing Open Data datasets, to identify best practice and to identify where further work is needed by the community. How can a dataset be evaluated to determine if it is "Open Data"? We will here apply the principles for "Open Data" according to the publication by (68). The FAIR principles are guidelines for the scientific community in developing Open Data datasets. FAIR is an acronym referring to the principles of Findable, Accessible, Interoperable and Reusable. This definition goes much further than publishing a dataset in location which is publicly available for download. The emphasis is on approaches to create datasets which can be understood and reused by third parties who were not involved with the original data collection. This involves considered use of metadata to describe the datasets and using formats which are both human and machine readable.

Secondly an approach is needed to determine if a dataset meets the FAIR principles. In this work, the approach taken by (69) in the paper "Are the FAIR principles fair?" is used. Here, a dataset is evaluated by classifying whether it complies with each of the categories of the FAIR principles according to one of the following options: "compliant"; "vague" and "not compliant". Each FAIR principal has a number of subcategories, that are shown in Table 1. The subcategories are defined in detail within the FAIR guidelines, together with examples of best practice datasets, which demonstrate how each subcategory can be successfully implemented. This work classifies the compliance of three occupant behavior Open Data datasets against each of the FAIR subcategories. The authors realize that this is a subjective approach and a different team of researchers could classify the same dataset in a different way. However, this subjective approach is still considered useful in providing general insights into which FAIR principles are being met by the research community and which FAIR principles are presenting a current challenge to Open Data authors.

Table 1 shows three occupant behavior datasets and their evaluation according to the FAIR principles. These datasets were chosen from the wider review of Open Data datasets as suitable examples to demonstrate the different approaches to data sharing and publishing. Here we discuss the findings for each FAIR Principle:

Findable: The Findable principle is essential, as for any dataset to be used it must first be found. This is largely a function of the repository on which the dataset is published. There are two overall principles here:

1. Globally unique and persistent identifiers should be used (F1). This means that the dataset should have an identifier, such as a persistent web address or URL, which is guaranteed both to be unique across the world and which will continue to exist into the foreseeable future. Researchers can then refer to a dataset using this identifier with confidence that there can be no confusion as to what is being referenced. Both Figshare and the UK Data Service (the hosts of the REFIT (70) and UK TUS datasets (72)) achieve this by assigning a Digital Object Identifier (DOI) to the dataset (for example the DOI of the REFIT dataset[14]). The OpenEI repository (which hosts the Langevin dataset (71)) does not ascribe a DOI or any equivalent, instead only a standard URL[15] is used which cannot be considered persistent and so they do not meet the F1 criterion. It should be noted that the FAIR F1 principle requires "every element of metadata and every concept/measurement in your dataset" to have a globally unique identifier. This is a challenging requirement as it requires elements of the dataset itself, such as the concept of "room air temperature" or "degrees Celsius" to have their own identifiers. None of the datasets under study achieve this and the REFIT and UK TUS datasets are rated as 'Vague' as a result.
2. Detailed metadata should be used (F2, F3 and F4). In this context, metadata refers to descriptors of the dataset such as the author list, the dataset identifier, a description of the dataset and keywords used to describe the data. Figshare and the UK Data Service meets these criteria and provide rich, structured metadata for the data items (Figshare and the FAIR data principles; [16]). OpenEI also provides metadata; however, this falls short of the standards required for FAIR. In particular, the Open EI fails on F3 as there is no unique identifier for the underlying dataset itself and there is only limited metadata provided (hence the 'Vague' rating for F2).

Accessible: The Accessible principle is largely concerned with the ability to access the dataset using standardised communication protocols. Examples of such protocols are http(s),

[14] https://doi.org/10.17028/rd.lboro.2070091.v1

[15] https://openei.org/datasets/dataset/one-year-behavior-environment-data-for-medium-office

[16] https://doi.org/10.6084/m9.figshare.7476428.v1

Table 1
An evaluation of occupant behavior Open Data datasets based on their compliance with the FAIR principles, and summary of existing Open Data datasets.

FAIR Principles			Datasets under study: REFIT (70)	Langevin (71)	UK TUS (72)
Findable	**F1.** (Meta)data are assigned a globally unique and persistent identifier		vague	not compliant	vague
	F2. Data are described with rich metadata (defined by R1 below)		compliant	vague	compliant
	F3. Metadata clearly and explicitly include the identifier of the data they describe		compliant	not compliant	compliant
	F4. (Meta)data are registered or indexed in a searchable resource		compliant	compliant	compliant
Accessible	**A1.** (Meta)data are retrievable by their identifier using a standardised communications protocol	**A1.1** The protocol is open, free, and universally implementable	compliant	compliant	compliant
		A1.2 The protocol allows for an authentication and authorisation procedure, where necessary	compliant	compliant	compliant
	A2. Metadata are accessible, even when the data are no longer available		compliant	vague	compliant
Interoperable	**I1.** (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.		vague	not compliant	vague
	I2. (Meta)data use vocabularies that follow FAIR principles		vague	not compliant	vague
	I3. (Meta)data include qualified references to other (meta)data		vague	not compliant	vague
Reusable	**R1.** Meta(data) are richly described with a plurality of accurate and relevant attributes	**R1.1.** (Meta)data are released with a clear and accessible data usage license	compliant	compliant	compliant
		R1.2. (Meta)data are associated with detailed provenance	vague	vague	compliant
		R1.3. (Meta)data meet domain-relevant community standards	vague	vague	compliant

ftp, email and telephone. The main point is that specialized software, which might need to be purchased, should not be required in order to access the datasets. The Accessible principle does not require the dataset to be publicly available, and datasets such as the UK TUS (which requires user registration before access) can still be compliant. In this case, Figshare, OpenEI and the UK Data Service are all repositories which meet the FAIR Accessible principle as they provide the datasets using the internet (i.e. the http(s) protocol). The OpenEI repository is rated as "Vague" for A2, as the policy on data persistence is unclear. The use of DOIs by Figshare and the UK Data Service offer a level of guarantee of data persistence and Figshare aims to guarantee "10 years of persistent availability"[17].

[17] https://knowledge.figshare.com/articles/item/how-persistent-is-my-research

Interoperable: The Interoperable principle is the most difficult of the FAIR principles to comply with. Interoperability refers directly to the dataset itself, rather than the repository on which is it hosted. Subcategory I1 refers to the need for "a formal, accessible, shared and broadly applicable language for knowledge representation" and provides the examples of RDF (Resource Description Framework), OWL (Web Ontology Language), DAML+OIL (a semantic markup language for Web resources) and JSON LD (a lightweight Linked Data format). These languages are the realm of the semantic web and the Linked Data movement. They are well established and formally specified in detail by the World Wide Web Consortium and other organizations They utilize globally unique identifiers as part of their data representation and place significant importance on the nature of the relationship between data points (not just the data

points themselves). It is beyond the scope of this work to describe these languages in detail but it is clear that using these languages would enable compliance with the Interoperable FAIR principle. For example, a well-formed OWL ontology that was developed by and shared amongst a research community could then be used as the schema to construct any number of RDF datasets. RDF datasets constructed in such a manner would be understandable by the research community and it would be a straightforward matter to link such datasets together. E.g. based on existing examples as presented in Section 6.

In the case of the datasets under study, the REFIT and UK TUS datasets are rated as "Vague" for the Interoperable criteria. Both datasets provide extensive information about the meaning of the variables recorded. The REFIT dataset utilizes a custom XML schema to provide meaning behind the responses and measurements, and the UK TUS provides detailed information about measurement variables and response options in Rich Text Format data dictionaries. Both these approaches are useful and other researchers have been able to reuse the datasets; however, neither approach utilizes a common, shared vocabulary or globally unique identifiers. The Langevin dataset provides information on variables and measurement responses in an Excel spreadsheet but with significantly less detail and structure. Due to this, the Langevin is rated as "Not compliant"; however, it should be noted that this has not prevented other researchers from also reusing this dataset.

Reusable: The Reusable principle refers to the ability of the dataset to be reused by others, rather than understood by other which is covered by Interoperability. R1.1 deals with the legal reusability and all three datasets are compliant as they provide suitable licenses for the datasets. R1.2 deals with providing a full description of the dataset for the purposes of reusing the data, i.e. the provenance of the dataset (rather than a description for assisting in finding the dataset, see the Findable principle). This includes, for example, the workflow that led to the data, how it was collected and how to cite the dataset. R1.3 states that the dataset should, if possible, be structured and presented in a style and format which is well-known and familiar (and preferable an established standard) with the relevant scientific domain. For R1.2 and R1.3 the UK TUS is considered compliant as there is significant accompanying documentation concerning the provenance of the dataset, and the dataset is presented according to the same standard (e.g., using SPPS files, RTF and tab-delimited files) as the many other social science datasets also available on the UK Data Service. The REFIT and Langevin datasets are rated as 'Vague' for these criteria as they provide only limited details about the dataset and how it was collected, and the data format approaches here are not community standards but more unique and limited.

9. Conclusion

This paper conducts a systematic review of current open data research and applications in terms of state-of-the-practice, infrastructure, metadata schema, privacy and ethics, and sharing principles. The review indicates the following challenges for sharing and utilizing open data sets: (i) lack of knowledge of available data sets; (ii) lack of detailed documentation on existing data sets; (iii) concerns on time consuming to provide open data and (iv) concerns on limiting release of the data. In addition, the review suggests that there are current solutions to address some of the challenges including: (i) methods for data anonymization methods that have been used in a number of studies to address data privacy issues; and (ii) data sharing principles, such as, FAIR exists but it is not widely used by researchers in the area of occupant behavior. Furthermore, the review of existing open data research and applications suggests taking a systematic view to open data, as it is more than just a set of technical solutions. For example, there might be a need to (i) consider the opportunity of separating data collection from the use of data for research due to the complexity associated with open data and its process, (ii) face a mix of open and non-open data from a data user's perspective that requires a different kind of data portal for managing data, (iii) need for policies and guidelines for protecting people who provide data, (iv) and sometimes perhaps need for a specific purpose of using open data in order to make an open data project successful. Thus, it requires to deal with many factors such as human, technical, and policy factors, which form an open data ecosystem including data owners and/or providers, data consumers, both open and protected data, repositories, applications platforms (e.g., for discovery, analysis, visualization, etc.), and policies and guidelines. Specifically, it appears that researchers in the building science community have the skills to do research using open data but the lack of structured and accessible open data may have prevented them from fully embracing the idea, hence hindering research in building science. While there are already many technical platforms and software tools that can assist researchers in creating, managing, and using open data, some key challenges still need to be addressed by the community collectively. Examples of such challenges include properly providing metadata on occupant data, ethics and privacy considerations in using building occupancy data, and the unfamiliarity with the best practice of sharing data such as the FAIR principles. Finally, a broad discussion about open data and outreach of best practice among stakeholders have to be encouraged in order to increase the awareness and the use of open data in the building science community.

10. Acknowledgments

This work has been performed within the framework of the International Energy Agency – Energy in Buildings and Communities Program (IEA-EBC) Annex 79 "Occupant-centric building design and operation". Bing Dong would like to thank the research support from the U.S. National Science Foundation CAREER Award (Award No. 1949372). Jens Hjort Schwee, Fisayo Caleb Sangogboye, and Mikkel Baun Kjærgaard would like to acknowledge funding by EUDP (Grant,

n. 64018-0558). Omid Ardakanian would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2019-04349]. Flora Salim would like to acknowledge the support of Humboldt Foundation and Bayer Foundation for her Humboldt-Bayer research fellowship, and Australian Research Council's funding (ARC LP150100246) for Nan Gao's PhD scholarship. Yimin Zhu would like to thank the research support from the U.S. National Science Foundation (Award No.: 1805914). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

[1] M. Jia, R. S. Srinivasan, A. A. Raheem, From occupancy to occupant behavior: An analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency, Renewable and Sustainable Energy Reviews 68 (2017) 525–540 (2017).

[2] B. Dong, D. Yan, Z. Li, Y. Jin, X. Feng, H. Fontenot, Modeling occupancy and behavior for better building design and operation—a critical review, Building Simulation 11 (5) (2010) 899–921 (2010).

[3] A. Mirakhorli, B. Dong, Occupancy behavior based model predictive control for building indoor climate—a critical review, Energy and Buildings 129 (2016) 499–513 (2016).

[4] Z. Yang, N. Li, B. Becerik-Gerber, M. Orosz, A systematic approach to occupancy modeling in ambient sensor-rich buildings, Simulation 90.8 (2014) 960–977 (2014).

[5] O. Ardakanian, A. Bhattacharya, D. Culler, Non-intrusive occupancy monitoring for energy conservation in commercial buildings, Energy and Buildings 179 (2018) 311–323 (2018).

[6] F. C. Sangogboye, K. Arendt, M. Jradi, C. Veje, M. B. Kjærgaard, B. N. Jørgensen, The impact of occupancy resolution on the accuracy of building energy performance simulation, in: Proceedings of the 5th Conference on Systems for Built Environments, 2018, pp. 103–106 (2018).

[7] S. Carlucci, G. Lobaccaro, Y. Li, E. C. Lucchino, R. Ramaci, The effect of spatial and temporal randomness of stochastically generated occupancy schedules on the energy performance of a multiresidential building, Energy and Buildings 127 (2016) 279 – 300 (2016).

[8] M. Jradi, K. Arendt, F. Sangogboye, C. Mattera, E. Markoska, M. Kjærgaard, C. Veje, B. Jørgensen, Obepme: An online building energy performance monitoring and evaluation tool to reduce energy performance gaps, Energy and Buildings 166 (2018) 196 – 209 (2018).

[9] I. B. Arief-Ang, M. Hamilton, F. D. Salim, Rup: Large room utilisation prediction with carbon dioxide sensor, Pervasive and Mobile Computing 46 (2018) 49–72 (2018).

[10] M. S. Rahaman, H. Pare, J. Liono, F. D. Salim, Y. Ren, J. Chan, S. Kudo, T. Rawling, A. Sinickas, Occuspace: Towards a robust occupancy prediction system for activity based workplace, in: 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), IEEE, 2019, pp. 415–418 (2019).

[11] F. C. Sangogboye, Data-driven methods for occupant sensing and privacy protection: With applications to enable smart and energy efficient buildings, Ph.D. thesis, University of Southern Denmark (2018).

[12] P. Murray-Rust, Open data in science, Ser. Rev. 34 (1) (2006) 52–64 (2006).

[13] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, DBpedia: a nucleus for a web of open data, in: K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux (Eds.), The Semantic Web, 2007, pp. 722–735 (2007).

[14] M. Janssen, Y. Charalabidis, A. Zuiderwijk, Benefits, adoption barriers and myths of open data and open government, Inf. Syst. Manag. 29 (4) (2012) 258–268 (2012).

[15] J. Gray, Towards a genealogy of open data, SSRN Electron (2015) 1–38 (September 2015).

[16] J. Tauberer, Open Government Data: The Book, 2nd ed., 2014 (2014).

[17] F. C. Sangogboye, R. Jia, T. Hong, C. Spanos, M. B. Kjærgaard, A framework for privacy-preserving data publishing with enhanced utility for cyber-physical systems, ACM Trans. Sen. Netw. 14 (3-4) (2018) Article 30 (2018).

[18] R. Jia, F. C. Sangogboye, T. Hong, C. Spanos, M. Baun Kjærgaard, Pad: Protecting anonymity in publishing building related datasets, in: Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys), 2017, pp. 1–10 (2017).

[19] J. Crusoe, U. Melin, Investigating open government data barriers, in: Electronic Government, 2018, pp. 169–183 (2018).

[20] D. Rudmark, A. Jordanius, Harnessing digital ecosystems through open data – diagnosing the swedish public transport industry, in: 27th European Conference on Information Systems (ECIS), 2019 (2019).

[21] S. Jansen, M. Cusumano, Defining software ecosystems: A survey of software platforms and business network governance, in: CEUR Workshop Proc., Vol. 879, 2012, p. 41–58 (2012).

[22] Z. Zhu, M. A. Wulder, D. P. Roy, C. E. Woodcock, M. C. Hansen, V. C. Radeloff, S. P. Healey, C. Schaaf, P. Hostert, P. Strobl, J.-F. Pekel, L. Lymburner, N. Pahlevan, T. A. Scambos, Benefits of the free and open landsat data policy, Remote Sens. Environ. 224 (February) (2019) 382–385 (2019).

[23] V. Gewin, An open mind on open data, Nature 529 (7584) (2016) 117–119 (2016).

[24] O. J. Reichman, M. B. Jones, M. P. Schildhauer, Challenges and opportunities of open data in ecology, Science 331 (6018) (2011) 703–705 (2011).

[25] X. Chen, S. Dallmeier-Tiessen, R. Dasler, S. Feger, P. Fokianos, J. B. Gonzalez, H. Hirvonsalo, D. Kousidis, A. Lavasa, S. Mele, D. R. Rodriguez, T. Simko, T. Smith, A. Trisovic, A. Trzcinska, I. Tsanaktsidis, M. Zimmermann, K. Cranmer, L. Heinrich, G. Watts, M. Hildreth, L. Lloret Iglesias, K. Lassila-Perini, S. Neubert, Open is not enough, Nat. Phys. 15 (2) (2019) 113–119 (2019).

[26] European Parliament and Council of the European Union, Regulations (EU) 2016/679 of the European Parliament and of the Council - general data protection regulation (GDPR), Official Journal of the European Union L119 (2016) 1—88 (2016).

[27] California State Legislature, California Consumer Privacy Act of 2018 (Jun. 2018).

[28] Office of the Australian Information Commissioner, Privacy Act 1988, Compilation No. 81 (Aug 2019).

[29] T. Dixon, J. Van de Wetering, M. Sexton, S.-L. Lu, D. Williams, D. Ulutas Duman, X. Chen, Smart cities, big data and the built environment: What's required?, Tech. rep., European Real Estate Society (ERES) (2016).

[30] 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 (2017).

[31] J. Roth, A. Bailey, S. Choudhary, R. K. Jain, Spatial and temporal modeling of urban building energy consumption using machine learning and open data, in: Sel. Pap. from ASCE Int. Conf. Comput. Civ. Eng., 2019, p. 459–467 (2019).

[32] S. Barker, A. Mishra, D. Irwin, E. Cecchet, P. Shenoy, J. Albrecht, Smart*: An open data set and tools for enabling research in sustainable homes, in: SustKDD, 2012, p. 6 (2012).

[33] B. Dong, M. B. Kjærgaard, M. De Simone, H. B. Gunay, W. O'Brien, D. Mora, J. Dziedzic, J. Zhao, Sensing and Data Acquisition, Springer International Publishing, 2018, Ch. 4, pp. 77–105 (2018).

[34] S. Pfenninger, J. DeCarolis, L. Hirth, S. Quoilin, I. Staffell, The importance of open data and software: Is energy research lagging behind?, Energy Policy 101 (2017) 211 – 215 (2017).

[35] N. Batra, J. Kelly, O. Parson, H. Dutta, W. Knottenbelt, A. Rogers,

A. Singh, M. Srivastava, Nilmtk: an open source toolkit for non-intrusive load monitoring, in: Proceedings of the 5th international conference on Future energy systems (e-Energy '14), 2014, pp. 265–276 (2014).

[36] G. Fierro, M. Pritoni, M. AbdelBaky, P. Raftery, T. Peffer, G. Thomson, D. E. Culler, Mortar: An open testbed for portable building analytics, in: Proceedings of the 5th ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys), 2018, pp. 172–181 (2018).

[37] T. Zhang, A. A. Zishan, O. Ardakanian, Odtoolkit: A toolkit for building occupancy detection, in: Proceedings of the Tenth ACM International Conference on Future Energy Systems (e-Energy '19), 2019, pp. 35–46 (2019).

[38] T. Hong, H. Sun, Y. Chen, S. C. Taylor-Lange, D. Yan, An occupant behavior modeling tool for co-simulation, Energy and Buildings 117 (2016) 272 – 281 (2016).

[39] B. Dong, K. Lam, Y. Huang, G. Dobbs, A comparative study of the ifc and gbxml informational infrastructures for data exchange in computational design support environments, in: Tenth International IBPSA Conference, 2007, pp. 1530–1537 (2007).

[40] M. Venugopal, C. Eastman, J. Teizer, An ontology-based analysis of the industry foundation class schema for building information model exchanges, Advanced Engineering Informatics 29 (4) (2015) 940–957 (2015).

[41] B. Dong, Z. O'Neill, Z. Li, A bim-enabled information infrastructure for building energy fault detection and diagnostics. automation in construction, Automation in Construction 44 (2014) 197–211 (2014).

[42] A. Mahdavi, M. Taheri, An ontology for building monitoring, Journal of Building Performance Simulation 10 (5-6) (2017) 499–508 (2017).

[43] A. Mahdavi, D. Wolosiuk, A building performance indicator ontology: Structure and applications, in: Proceedings of Building Simulation 2019: 16th Conference of IBPSA, 2019, pp. 385–390 (2019).

[44] B. Balaji, A. Bhattacharya, G. Fierro, J. Gao, J. Gluck, D. Hong, A. Johansen, J. Koh, J. Ploennigs, Y. Agarwal, M. Bergés, D. Culler, R. K. Gupta, M. B. Kjærgaard, M. Srivastava, K. Whitehouse, Brick : Metadata schema for portable smart building applications, Applied Energy 226 (2018) 1273 – 1292 (2018).

[45] T. Hong, S. D'Oca, W. J. Turner, S. C. Taylor-Lange, An ontology to represent energy-related occupant behavior in buildings. part i: Introduction to the dnas framework, Building and Environment 92 (2015) 764 – 777 (2015).

[46] W. J. Turner, T. Hong, A technical framework to describe energy-related occupant behavior in buildings, in: Proceedings of the Behavior, Energy & Climate Change conference, BECC 2013, 2013 (2013).

[47] T. Hong, S. D'Oca, S. C. Taylor-Lange, W. J. Turner, Y. Chen, S. P. Corgnati, An ontology to represent energy-related occupant behavior in buildings. part ii: Implementation of the dnas framework using an xml schema, Building and Environment 94 (2015) 196 – 205 (2015).

[48] Z. D. Belafi, T. Hong, A. Reith, A library of building occupant behaviour models represented in a standardised schema, Energy Efficiency 12 (3) (2019) 637–651 (2019).

[49] H. Lange, A. Johansen, M. B. Kjærgaard, Evaluation of the opportunities and limitations of using IFC models as source of building metadata, in: Proceedings of the 5th Conference on Systems for Built Environments, BuildSys 2018, 2018, pp. 21–24 (2018).

[50] C. Thornton, F. Hutter, H. H. Hoos, K. Leyton-Brown, Auto-weka: Combined selection and hyperparameter optimization of classification algorithms, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp. 847–855 (2013).

[51] J. C. Chang, S. Amershi, E. Kamar, Revolt: Collaborative crowdsourcing for labeling machine learning datasets, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 2334–2346 (2017).

[52] M. Kaur, F. D. Salim, Y. Ren, J. Chan, M. Tomko, M. Sanderson, Shopping intent recognition and location prediction from cyber-physical activities via wi-fi logs, in: Proceedings of the 5th Conference on Systems for Built Environments, 2018, pp. 130–139 (2018).

[53] M. Pham, S. Alse, C. A. Knoblock, P. Szekely, Semantic labeling: a domain-independent approach, in: International Semantic Web Conference, 2016, pp. 446–462 (2016).

[54] J. Koh, D. Hong, R. E. Gupta, K. Whitehouse, H. Wang, Y. Agarwal, Plaster: an integration, benchmark, and development framework for metadata normalization methods, in: Proceedings of the 5th Conference on Systems for Built Environments, BuildSys 2018, 2018, pp. 1–10 (2018).

[55] W. Wang, Z. Chen, J. Liu, Q. Qi, Z. Zhao, User-based collaborative filtering on cross domain by tag transfer learning, in: Proceedings of the 1st International Workshop on Cross Domain Knowledge Discovery in Web and Social Network Mining, 2012, pp. 10–17 (2012).

[56] The Australian National Data Service (ANDS), Data sharing considerations for human research ethics committees (2018).

[57] D. Yan, T. Hong, B. Dong, A. Mahdavi, S. D'Oca, I. Gaetani, X. Feng, Iea ebc annex 66: Definition and simulation of occupant behavior in buildings, Energy and Buildings 156 (2017) 258–270 (2017).

[58] National Health and Medical Research Council, ational statement on ethical conduct in human research, National Health and Medical Research Council (2007).

[59] E. T. Slonecker, D. M. Shaw, T. M. Lillesand, Emerging legal and ethical issues in advanced remote sensing technology, Photogrammetric engineering and remote sensing 64 (6) (1998) 589–595 (1998).

[60] K. Crawford, M. Whittaker, M. C. Elish, S. Barocas, A. Plasek, K. Ferryman, The ai now report: The social and economic implications of artificial intelligence technologies in the near-term, AI Now public symposium, hosted by the White House and New York University's Information Law Institute (2016).

[61] L. Sweeney, k-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (05) (2002) 557–570 (2002).

[62] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, L-diversity: privacy beyond k-anonymity, in: ICDE'06, 2006, p. 24–24 (2006).

[63] C. Dwork, Differential privacy, in: Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II (ICALP'06), 2006, p. 1–12 (2006).

[64] I. Kotsogiannis, A. Machanavajjhala, M. Hay, G. Miklau, Pythia: Data dependent differentially private algorithm selection, in: Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD '17), 2017, p. 1323–1337 (2017).

[65] M. Malekzadeh, R. G. Clegg, H. Haddadi, Replacement autoencoder: A privacy-preserving algorithm for sensory data analysis, in: 2018 IEEE/ACM Third International Conference on Internet-of-Things Design and Implementation (IoTDI), 2018, p. 165–176 (2018).

[66] L. Rocher, J. M. Hendrickx, Y.-A. De Montjoye, Estimating the success of re-identifications in incomplete datasets using generative models, Nature Communications 10 (1) (2019) 3069 (2019).

[67] J. H. Schwee, F. C. Sangogboye, M. B. Kjærgaard, Evaluating practical privacy attacks for building data anonymized by standard methods, in: IoTSec '19, 2019 (2019).

[68] M. Wilkinson, M. Dumontier, I. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, L. da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. Evelo, R. Finkers, A. Gonzalez-Beltran, A. Gray, P. Groth, C. Goble, J. Grethe, J. Heringa, P. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. Lusher, M. Martone, A. Mons, A. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. Swertz, M. Thompson, J. Van Der Lei, E. Van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016).

[69] A. Dunning, M. De Smaele, J. Böhmer, Are the fair data principles fair?, International Journal of digital curation 12 (2) (2017) 177–195 (2017).

[70] S. Firth, T. Kane, V. Dimitriou, T. Hassan, F. Fouchal, M. Coleman, L. Webb, REFIT Smart Home dataset (6 2017).

[71] J. Langevin, P. L. Gurian, J. Wen, Tracking the human-building interaction: A longitudinal field study of occupant behavior in air-conditioned offices, Journal of Environmental Psychology 42 (2015) 94–115 (2015).

[72] J. Gershuny, O. Sullivan, United kingdom time use survey, 2014-2015 (2017).