Ontology-Based Modeling of Privacy Vulnerabilities for Data Sharing

Jens Hjort Schwee, Fisayo Caleb Sangogboye, Aslak Johansen, Mikkel Baun Kjærgaard

Ontology-based Modeling of Privacy Vulnerabilities for Data Sharing

Jens Hjort Schwee[1 [0000−0001−9176−2024]], Fisayo Caleb Sangogboye[1 [0000−0001−9995−758X]], Aslak Johansen[1 [0000−0001−6133−9030]], and Mikkel Baun Kjærgaard[1 [0000−0001−5124−744X]]

University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark
{jehs,fsan,asjo,mbkj}@mmmi.sdu.dk

Abstract. When several parties want to share sensor-based datasets it can be difficult to know exactly what kinds of information can be extracted from the shared data. This is because many types of sensor data can be used to estimate indirect information, e.g., in smart buildings a CO_2 stream can be used to estimate the presence and number of occupants in each room. If a data publisher does not consider these transformations of data their privacy protection of the data might be problematic. It currently requires a manual inspection by a knowledge expert of each dataset to identify possible privacy vulnerabilities for estimating indirect information. This manual process does not scale with the increasing availability of data due to the general lack of experts and the associated cost with their work. To improve this process, we propose a privacy vulnerability ontology that helps highlight the specific privacy challenges that can emerge when sharing a dataset. The ontology is intended to model data transformations, privacy attacks, and privacy risks regarding data streams. In the paper, we have used the ontology for modeling the findings of eight papers in the smart building domain. Furthermore, the ontology is applied to a case study scenario using a published dataset. The results show that the ontology can be used to highlight privacy risks in datasets.

Keywords: Open Data, Data Anonymization, Modeling Methodologies, Data Publishing, Data Privacy, Privacy-Preserving Data Publishing

1 Introduction

Open data has a great potential for improving scientific practices in terms of transparency and efficiency [23]. Data is also the foundation for new intelligent data-driven software solutions in many application areas. The large sensor-networks installed in a modern smart building can monitor almost every aspect of the building and the occupants inside [15]. This includes data about how the building is being used by occupants and control systems. e.g., from occupancy counting sensors and the building's building management system (BMS). Sharing such information with contractors who perform regular tasks – such as catering,

cleaning, and facility management – would enable them to develop data-driven applications for these operations. Considering a scenario where a catering company knows how many people are actually on the premises, they would know exactly how many to cook for. If they also know who these people are, then they could have access to dietary needs and cook according to preferences.

The European Data Portal has made a guideline for how to create a strategy for sharing open data [8], which consists of:

1. ***Ambition*** The strategy starts with setting the ambition for publishing data, including collecting a clear picture of the current data sharing situation in the organization. Furthermore, defining the intended situation.
2. ***Strategy*** Create a strategy for an open data policy, which among others, includes identifying all data in the organization, aligning the legal aspects, and formalizing key performance indicators to measure the progress, both for sharing the data and the impact of sharing it.
3. ***Policy*** Define the policy benefits for sharing the data. This includes defining, the scope, the goals, and the data types and quality of the data. Finally, the legal aspects, which include licensing, intellectual property, and privacy aspects.

Implementing such a strategy requires finding an appropriate data platform for the data, and making the data understandable for the recipient, as well as identifying the expected use from the open data community. Finally, a plan for how to maintain the data should be made, such that data release is up to date.

This process includes several stakeholders: The management of the company who develops the open data strategy; the data publisher, who is the one processing the data for sharing and who also is doing the privacy assessment, as well as maintaining the data; the data providers who either have actively participating in data collection (e.g., notes and sensor data from a smart watch) or passively being observed (e.g., data from a building-wide video analytics system); the data recipient, who are using the shared data.

When sharing data a data publisher needs to consider the privacy implications of sharing data. This is both to comply with privacy laws and regulations and to respect the interests of the data providers. One of the many challenges when sharing data is to identify the potential privacy implications of each part of the dataset, as well as combinations with available information through side channels. Combinations with other data can be the base of a privacy attack on the data which result in personal information about the data provider being revealed. Even more difficult is identifying privacy attacks, which can be achieved using artificial intelligence (AI) methods, as the AI area is a continuously moving target due to advancements in research. This makes it very difficult to keep an updated picture of the potential privacy risks. Here anonymizing the data and deleting any information related to the participants (e.g., name and height) is a method for privacy protection to limit such attacks on the data, but the current method also has their shortcomings. State-of-the-practice (SoP) methods for anonymization have, in recent research [26, 28] been found insufficient to protect the released data. Indicating that the data publisher did not know

the specific inference and data linkage possibilities, as well as the privacy risks related to the data release. Therefore, there is a need for solutions to address these problems.

Consider an example from the smart building domain, where a data publisher releases CO_2 data collected in a single room. The data publisher performs a privacy assessment for the CO_2 stream and identifies some privacy risks for the stream. Based on these, an anonymization strategy is selected for the data release. However, the data publisher needs also to be aware of the inference and data linkage possibilities using the stream, e.g., in the case of a CO_2 data stream an AI-based transformation model can estimate the number of occupants in the monitored area [4]. The combination of such inference possibilities and lack of privacy risks knowledge can lead to the data not being adequately protected for the data release. Furthermore, a data publisher must consider laws like the European General Data Protection Regulation (GDPR) [11], the EU's ePrivacy Directive [22], the California Consumer Privacy Act (CCPA) [7], and the Australian Privacy Principles (APPs) [20], which amongst others, defines the personally identifiable information that organizations are allowed to store and share, without reasonable cause or user consent. Many smart buildings are publicly accessible buildings. Therefore, an attacker can physically observe the building and combine ground truth with the published data to potentially infer complete knowledge about the building's use and conditions. This makes it challenging to select and create anonymization methods for smart building data.

In this paper, we address the very important privacy assessment step of the sharing process. The problem is that it can be very difficult for a data publisher to identify the potential privacy implications for a specific dataset. The data publisher might not have a full overview of the privacy risks related to each piece of data. This can lead to sharing of privacy problematic data which is a problem both for the individual data providers and for the publisher in terms of laws and regulations. Knowledge of the potential privacy risks is thus critical. Therefore, we present an ontology that for a given dataset can model privacy risks, privacy attacks, and possible data transformations. Previous work document a long range of possible transformations that can be applied to smart building datasets. However, it is impossible for data publishers to be experts and be up to date with the latest knowledge. In addition, the data publishers also need to consider that several transformations might be combined to reveal a certain piece of information from a dataset. Therefore, there is a need for a common format for modeling privacy risks. From a top-level view, the proposed ontology is illustrated in Figure 1, including elements for data, transformations, privacy risks, and attacks, and the relationships among them. Using this ontology, the theoretical privacy risks associated with sharing some data can be modeled. The model can be used to consider privacy risks before sharing data. The paper will focus on the smart building domain, and will, therefore, propose a solution targeted this domain. However, the solution has a clear potential to be extended to other domains.

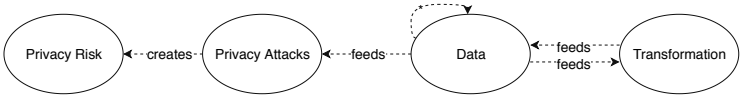

Fig. 1. The elements in the ontology. Includes classes and object relations.

The rest of the paper is structured as follows: Section 2 covers related work; section 3 presents an analysis of a set of papers within the smart building domain to identify the concepts classes to be included in the ontology; section 4 presents the design of the ontology; section 5 presents an instance example of the ontology for each of the analyzed papers; section 6 a case study combines all the individual models into a single instance of the ontology, which considers the privacy risks for a specific data release. Finally, section 7 and 8 we will discuss and conclude on the paper.

2 Related Work

This section covers the related work including, including prominent privacy-preserving methods, e.g., methods for Privacy-Preserving Data Publishing (PPDP). The section also covers data transformations that can be performed on smart building data. Furthermore, the section also studies existing smart building ontologies, as well as the components necessary for creating instances of it and taxonomies for the sensitivity of attributes.

2.1 Protection Methods

A data publisher must consider if there is a need for protecting sensitive information in a dataset or the identity of monitored individuals using privacy-preserving data mining or PPDP methods, respectively [25]. The field of PPDP has developed several privacy models. Some of the more prominent ones include k-anonymity [30], l-diversity [18], and δ-presence [19] differential privacy [9]. These models protect against the privacy attacks of record-linkage, attribute-linkage, table-linkage, and probabilistic attacks, respectively.

More recently, Pappachan et al. [21], have developed a framework which can be used for configuring privacy settings of users and enforcing these when collecting and sharing user data in the smart building domain. The framework contains three elements: Internet of Things (IoT) resource registries, IoT assistants, and privacy-aware smart buildings. Resource registries are a collection of policies and sharing practices of IoT technologies. IoT assistants notify users about the policies of the IoT devices. Furthermore, the framework provides configuration for privacy settings, which enables the users to define which information they are willing to share with specific services. The operator of a smart building can also define which services are being used in the building, which can override the users' privacy settings. The privacy-aware smart building receives the privacy settings and uses them when collecting user data and sharing data.

Schwee et al [28], have explored state-of-the-practice (SoP) PPDP methods on a published building dataset. They found that SoP methods like suppression

were insufficient to protect the identity of rooms in the published dataset. As a result of this, identifying a potential privacy risk for such building data showing that occupancy data and room schedules can reveal the identity of a room.

2.2 Data Transformation

The recent advancements in AI have enabled data-driven creation of models that can by estimation transform a combination of input data to relevant information. A number of such transformation has been developed within the field of smart buildings: Arief-Ang et al. [4], have demonstrated that using nothing but CO_2 measurements, the amount of occupants in a room can be estimated. Furthermore, [3], presented how to generalize this method using domain-adaptation. Ardakanian et al. [1], have proposed an occupancy detection technique using fine-grained measurements from Variable Air Volume (VAV) systems. The results show that the estimated occupancy patterns can be used for per-zone schedules. Hence it can be used to estimate occupancy. In Section 3 more examples of such transformations are given when analyzing specific examples.

2.3 Ontologies and Taxonomies

Ontologies enable the structuring data from data collection and also to structure the description of the data collection. A number of ontologies have been proposed within the smart building field including Haystack, Industry Foundation Classes (IFC), and BRICK to describe the building and the installed data collection infrastructure. A prominent example is BRICK [5] which is an ontology to structure data about smart buildings. BRICK structures data based on the Resource Description Framework (RDF) [17] to define the relations between the elements (e.g. system or building elements) in the ontology as well as hierarchies of such elements. Having a BRICK model of a smart building enables the use of SPARQL [24] queries to discover control systems and the location of sensors. This enables applications to use the model to discover resources instead of hardcoding an application to the specific devices in a building. Furthermore, the model can be used to identify the available sensors at each location and explore how they are colocated.

Within the privacy field, a taxonomy for privacy sensitivity attributes can be found in [12]. It defines four types of attributes: Explicit identifiers, quasi-identifiers, sensitive attributes, and non-sensitive attributes. The explicit identifiers identify a record owner. The quasi-identifiers can, as a set, potentially identify a record owner. The sensitive attributes contain attributes which are person-specific information. The set of non-sensitive attributes is all attributes that do not fall into any of the others. The attributes in the Explicit identifier, quasi identifier, and sensitive attributes, need to be protected before they can be shared.

2.4 Summary

The mentioned privacy protection methods can be applied when it is known which potential privacy risks a dataset needs to be protected against. However, a data publisher has to identify the risks of the data, the inference opportunities, and the data linkage possibilities. The mentioned ontologies only model the physical relationships in the building, and thus leaves a gap when it comes to modeling data-sharing challenges. The privacy attribute taxonomy is in its current form difficult to apply on time-series data, which a lot of smart building data are. This is because the privacy risks for individuals often occur in a non-statically manner, e.g., outlines in the data streams. Therefore, to help data publishers there is a need for means for modeling privacy risks that incorporate knowledge created by the scientific community. Such models can, among others support data publishers in their privacy assessments for a specific dataset.

In this paper, we propose a method for modeling data, privacy risks, transformations, and privacy attacks and their relationships in a graph-based model using Resource Description Framework (RDF) triples. The graphs can be used for highlighting the potential privacy risk for a specific dataset before sharing the data. Thereby, giving the data publisher a mean to structure knowledge about the potential privacy risks and the ability to make better-informed decisions before sharing data.

3 Analysis of Domain Cases

In this section, we analyze existing work to identify the elements needed in an ontology for modeling privacy risks, privacy attacks, and possible data transformations. Six papers were selected to cover many types of sensor data, data transformations, and extracted information. Furthermore, all of them have been recently published in major conferences or journals in the field.

Electricity consumption measured by smart meters is a widely used sensor modality in buildings to capture electricity use and related occupant behavior. As an example, Sonta et al. [29], have proposed methods that for each desk as input use electricity consumption data for all available plugs. Their method can transform this data into desk-level activities of the occupant. In addition, the method can also produce a social network for the social interactions between the occupants in a whole office. Another example is Kleiminger et al. [16] that propose a method that from electricity consumption measured at a household granularity can estimate household occupancy. Furthermore, Beckel et al. [6] have created a model that from electricity consumption data at the household level can estimate several demographic parameters including the age of the residents, marital status, employment and the amount of bedrooms.

Door monitoring has been studied by many authors to capture room movement in buildings. As an example, Khalil et al. [14] propose a method that as input uses data from ultrasonic-based distance sensors in each door. Their method can from these measurements identify occupants passing through the

door. Furthermore, the authors have proposed a method for detecting if the occupant passing the door is using a phone, holding a handbag, or a backpack.

Another relevant type of information is room occupation and number of occupants. As an example, Kjærgaard et al. [15] propose methods that from passive infrared movement or video-based sensors can estimate the presence and number of occupants. The methods can also via machine learning predict the future amount of occupants in an area. The spatial resolution of the predictions is similar to the monitored resolution, e.g., if monitoring private offices, then the prediction will be on the private office level. Likewise, Sangogboye et al. [27] have proposed a method that from passive infrared movement, temperature, and CO_2 sensor streams estimates the future amount of occupants in an area.

In a shopping context knowledge about the movement of shoppers and their intent is highly relevant to optimize shops. As an example, Kaur et al. [13] have proposed a method that from Wi-Fi logs estimates shoppers' physical movement. This physical information is then combined with shoppers' cyberbehvior based on weblogs to estimate the intent of the individual shoppers.

To study humans' health and wellbeing in indoor environments their activities and metabolic rate are too highly relevant types of information. As an example, Dziedzic et al [10] have proposed an estimation model using the movement skeleton joints of the human body captured by a Microsoft Kinect installed in a household. The skeleton movement data is used to estimate an occupant's metabolic rate and activities. Furthermore, the paper highlights that using the data in conjunction with external data, one can estimate the weight of the occupant.

With the methods presented in the covered papers, one can extract several types of information from building-related data. An overview of what we have found in these analyses can be seen in Table 1. The table highlights which kind of information each of the methods in the papers needs as input to extract the specify type of information. Furthermore, the table highlights that the methods can extract many different types of information from building-related data. The papers use data in the form of time-series, graph, external, and metadata which an ontology must have classes to model.

4 Ontology-based Modeling

In this section, we propose an ontology for modeling potential privacy attacks, privacy risks, and transformations. The ontology was developed based on our analyze results of existing work.

The analysis identified a number of different data types, namely: time-series, external, graph, and metadata. The analysis also highlighted a number of transformations that also need a concept in the ontology. Furthermore, the ontology has to model possible privacy risks and privacy attacks. All of the concepts are to be modeled using RDF. An overview of the ontology and the concepts can be found in Figure 1. This gives the following concepts in the ontology:

- ***Data*** superclass, which is the superclass of all data types.

Table 1. Overview over the papers analysed, and what information can be learned by the methods of each of the papers.

Paper	Sensor Modality	Necessary resolution			Reveal Information of type						
		Id	Spatial	Temporal	Relations	Behavior	Actions	Intent	Demographics	Health	Id
Sonta et al. [29]	Electricity Consumption			X	X						
Khalil et al. [14]	Door Distances	X	X			X					X
Kjærgaard et al. [15]	Door Openings	X	X			X					
Kjærgaard et al. [15]	Occupant counts	X	X			X					X
Kjærgaard et al. [15]	PIR movement	X	X			X					X
Sangogboye et al. [27]	PIR movement & Temperature & CO_2	X	X			X					
Kaur et al. [13]	Web query Log & WiFi access-point association log & Shopping mall layout	X	X			X		X			
Kleiminger et al. [16]	Electricity consumption	X	X			X					
Beckel et al. [6]	Electricity consumption			X					X		
Dziedzic et al. [10]	Body skeleton joints	X	X				X	X		X	

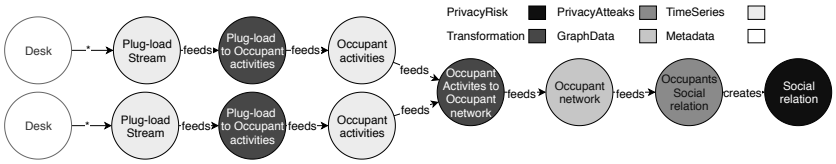

Fig. 2. Instance of the ontology, using the method presented by Sonta et al. [29].

- ***TimeSeries*** model any time-series data in relation to the data release.
- ***External*** any external data which might be available in relation to the monitored area.
- ***Metadata*** all relevant associated data that can be available about the data or the monitored context.
- ***Graph*** is representing graphs data.

– ***Transformation*** is a concept class that models the transformation and inference possibilities, which can be performed on the data in the model. The class must have one or more feed relation from Data classes which represent the relations that are involved in providing input data to the transformation. Furthermore, the class has an outgoing feed relation to other Data classes to represent the output of transformations in terms of new types of data.
– ***PrivacyAttack*** is a class that models the potential attacks which can be performed on Data sources, which can be modeled using the feed relation. The attacks may only be executed if all of its inputs are available. The class can have relations to PrivacyRisk classes using a create relation.
– ***PrivacyRisk*** represents a potential privacy risk in relation to the data release. We use the term privacy risk to highlight that only if an attack is performed it results in an actual leak until then there is a risk that this might happen.

As an example, using the ontology, transformations, privacy attacks, and risk of the method by Sonta et al. [29] is shown in Figure 2. We have modeled the context monitored as metadata and only two desks for clarity of the example. The streams of plug-load consumption measurements have been modeled as TimeSeries classes, these streams can be transformed into a time-series of occupant activities, using the proposed transformation. This is modeled as a transformation and produces a new TimeSeries. Based on these occupant activities, there is another transformation that can construct an occupant network, which can be represented by the Graph class. Finally, using the network, there can be performed PrivacyAttack estimating social relations among the occupants of the desks which is modeled as a PrivacyRisk. The instance of the ontology can be used to identify that if releasing plug-load consumption data, which is monitored on a desk level, this can potentially reveal social relations among the occupants.

5 Models of Domain Cases

To illustrate the applicability of the ontology we apply it in a number of cases. The goal is to be able to model the different cases with the elements of the

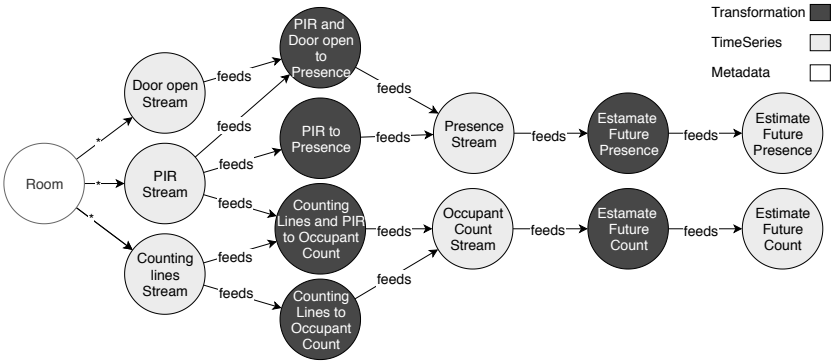

Fig. 3. Instance of the ontology, which models the findings of the paper Kjærgaard et al. [15]. We have chosen to model a single room as the monitored area.

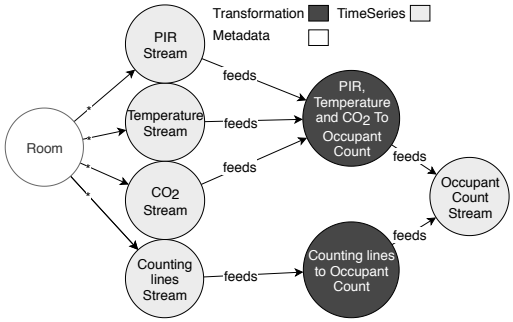

Fig. 4. Instance of the ontology, which models the findings of the paper Sangogboye et al. [27]. We have chosen to model a single room as the monitored area.

proposed ontology. The ontology is applied to model the transformations and privacy risks identified in each of the case papers. In the models, we will only include elements that have been mentioned by the papers. Therefore, we do not include any potential other privacy risks or privacy attacks which could, in theory, be performed upon the data.

As the first case we consider the transformation methods described by Kjærgaard et al. [15]. We have modeled all of them in a single instance of the ontology, which can be found in Figure 3. Using the model, it can be observed that the three time-series streams for CO_2, door opening measurements, and vision-based sensor count lines can be used to estimate the future presence and amount of occupants in the monitored zone.

The instance for Sangogboye et al. [27], can be found in Figure 4. The model highlights the possibility of estimating occupant counts using a combination of Passive infrared sensor (PIR) movement, CO_2, and temperature streams, in a monitored area.

The information found in Khalil et al. [14], has been modeled in the instance of the ontology found in Figure 5. The model includes the privacy attack and risks on the user identity, which has been modeled as a potential privacy attack. The attack can be performed using the time-series streams about user behavior.

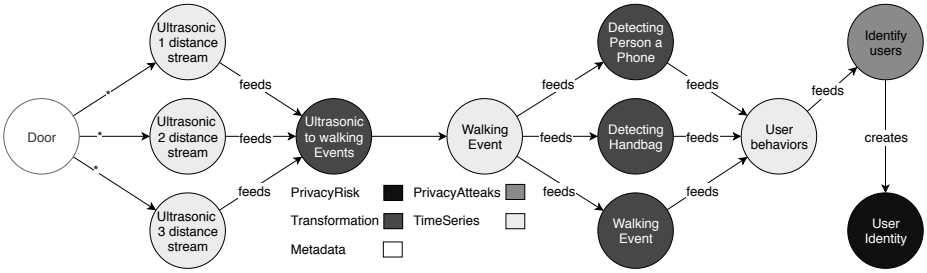

Fig. 5. Instance of the ontology, using the findings of the paper Khalil et al. [14].

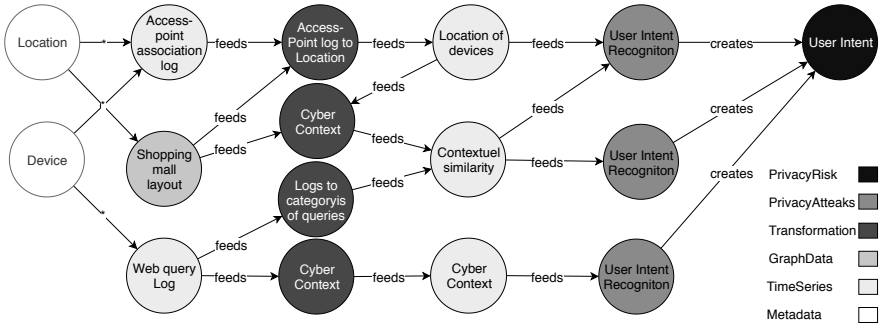

Fig. 6. Instance of the ontology, which models the findings of the paper Kaur et al. [13].

Using the model it can be observed that based on the three distance sensors installed in a single door, it can be used to identify user behavior and in some cases identify the occupant passing through the door.

Figure 6 shows the instance of the ontology with the information found in [13]. The model includes the privacy attack and risks on the user intent, which was modeled as a potential privacy risk. The information can be exposed using three different privacy attacks, captured in the model.

The instance for the paper [16], can be seen in Figure 7. This instance highlights that several methods can be used to detect occupancy presence using electricity consumption data from smart meters.

The instance for the paper [6], can be seen in Figure 8. We have only modeled a part of the privacy risks for simplicity. In the paper, they have proposed two methods for the detection of household properties for the occupants living there, namely a regression model and a classification model. We have modeled each of them as a privacy attack with associated privacy risks.

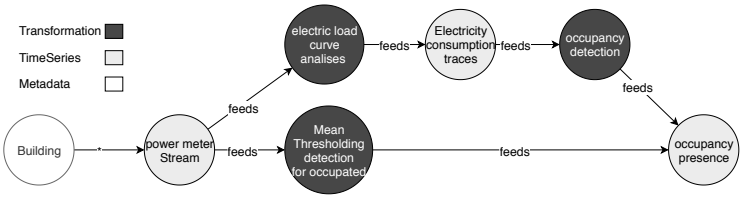

Fig. 7. Instance of the ontology, using the findings of the paper Kleiminger et al. [16].

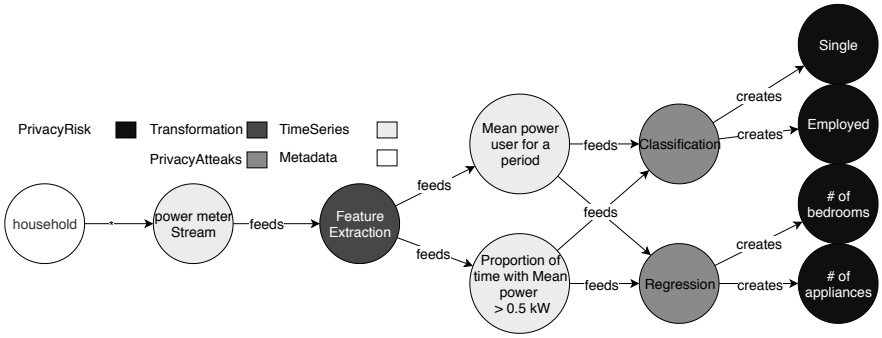

Fig. 8. Instance of the ontology, which models the findings of the paper Beckel et al. [6].

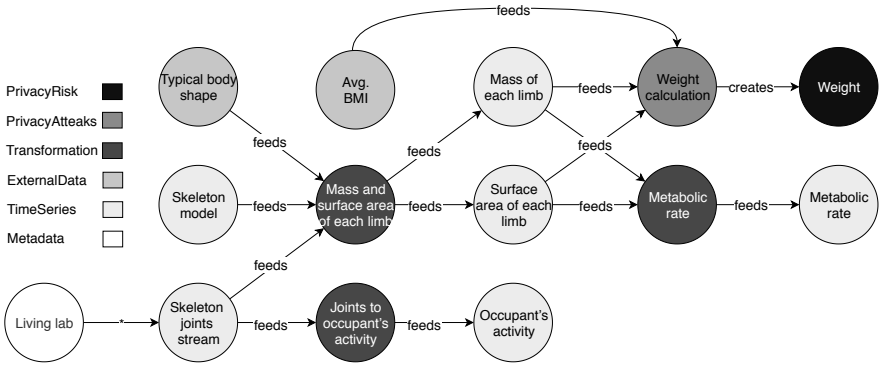

Fig. 9. Instance of the ontology, using the findings of the paper Dziedzic et al. [10].

The instance for the paper [10], can be seen in Figure 9. In this instance, it can be observed that weight and metabolic rates for the monitored individual can be estimated by fusing body skeleton joint data with external data. Furthermore, the joint data can be used to determine the occupant's activities.

As shown the ontology was able to model the elements of the particular domain cases. This highlights the potential of the ontology. However, as discussed later the ontology has some limitations which should be explored in future work.

6 Modeling an Open Dataset

In this section, we present a model capturing expert knowledge from several papers [1, 3, 4] and apply it for privacy assessment for an open dataset. The model will combine the findings of the papers into a single instance of the ontology. We use the model to identify privacy risks for a real-world dataset already shared as open data [2]. The papers include the focus on data from ventilation systems in the form of variable air volume (VAV) data which captures the inflow of air into a room and CO_2 which negatively correlate with oxygen content and thereby represent air quality. The papers propose methods to map these data sources to estimate occupant presence. This can, in the ontology, be directly modeled as a transformation from each of the data sources to occupancy presence.

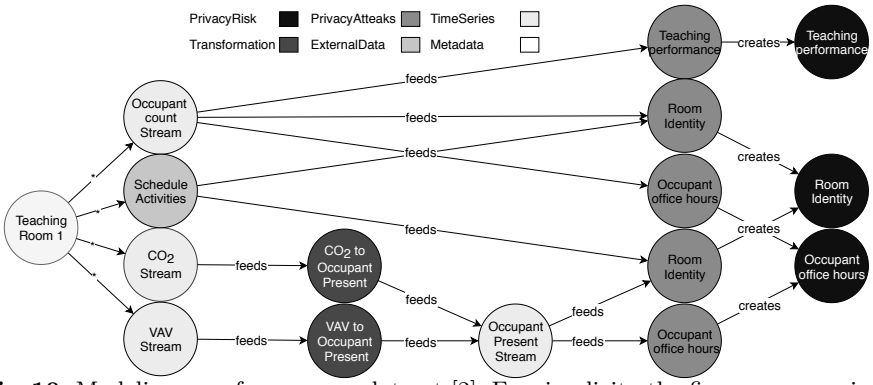

Fig. 10. Modeling case for an open dataset [2]. For simplicity the figure cover a single room from the released dataset with associated possible data transformations, privacy attacks, and privacy risks.

In terms of privacy attacks on occupant presence data, [28] highlighted several types of possible attacks. Firstly, they found that the room identity potentially can be identified using occupancy presence and counts combined with scheduled activities. Secondly, the occupancy presence and counts could also be used for identifying the working hours of an occupant working in the monitored area. Finally, it was found that the occupant counts for a teaching room, can be used to estimate teaching performance for each lecture in the facilities. These privacy attacks create three associated privacy risks. The model is shown in Figure 10.

As a dataset for the case study, we have used [2]. We consider data from a single room with the following sensors: CO_2, VAV, and amount of occupants. These relations are modeled in Figure 10. Here the metadata about room context and published data sources have been linked with the identified transformations, as well as the information about the room schedules that are publicly available [28].

With this model, a data publisher can for each type of data follow the graph and see what risks each data type result in. For instance, a data publisher can use the model to identify that occupant counts can potentially be used to estimate teaching performance and occupant office hours. The data publisher can also go the other way and from a privacy, risk backtrack what types of data result in this privacy risk. For instance, with the model one can observe that if the identity of a teaching room is to be hidden, there is a need for looking into how to anonymize the counts' data, as well as the presence data.

7 Discussion

The proposed ontology can model theoretical privacy risks in terms of how they relate to different data types. Therefore, it does not consider effects, such as the accuracy of data or the increasing level of uncertainty when data is transformed multiple times. However, since the ontology was designed to model the theoretical risks, this was not considered.

While completing the instances in the case study section, it was found that some of the findings for each paper depended on sensor deployment, e.g., in [15].

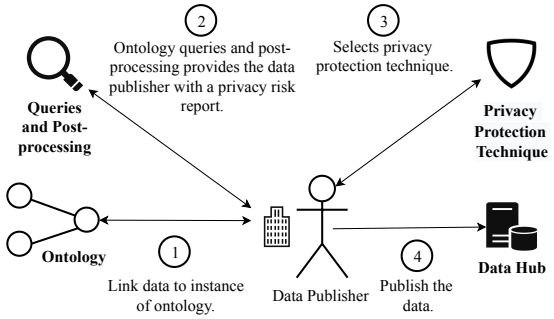

Fig. 11. The process for using the envisioned ontology for data publishing

This indicates that the current version of the ontology might need a concept that can be used to model a monitored context, which would be a relevant element to explore in future work. This might also help to combine the findings in the papers, hereby modeling the privacy risks to a specific context. Furthermore, in some of the models, it can be observed that the data can be transformed into a number of time-series, which in some cases might pose a privacy risk, e.g., Figure 9, where the data can be used to estimate occupant's activities or metabolic rates. However, this could be addressed by adding a privacy attack and privacy risk to each of them.

In this paper, we propose a privacy risk mapping ontology to improve the privacy assessment process for data publishing. However, we are aware that the ontology cannot stand on its own and would be hard for a non-domain expert to use. Therefore the ontology is only a piece of an envisioned tool-chain, which is sketched in Figure 11. The process is as follows: 1) the data publisher maps the ontology instance, which has the privacy risks and attacks modeled, with the data considered for data publishing; 2) the ontology's queries and post-processing provide the data publisher with results of the potential inference, vulnerabilities, and attack vectors for each of the monitored contexts; 3) based on the output of the former step the data publisher selects a privacy protection technique; and 4) the final data is released. In this tool-chain, the ontology is to be used as a foundation and the queries and post-processing are to be reused between implementations for different datasets. The intended output of the tool-chain is privacy risks and possible methods for how to protect the dataset.

8 Conclusion

We set out to design an ontology enabling data publishers to make preemptive privacy assessments before sharing open datasets. To design the ontology we analyzed eight recent papers within the field of smart buildings to distill examples of data transformations and privacy risks. The papers cover many different forms of sensor modalities and how they can be transformed into information like occupant identity, actions, behaviors, and intents. The proposed ontology models data transformations, and associated privacy attacks and risks. We evaluated the ontology by constructing individual models for the methods of the eight papers. The result was that in all cases we could describe the datasets and methods of

a paper by an instance of the ontology. Furthermore, we created a larger model based on related work which was applied to a particular dataset to be released. To support a privacy assessment the model could for this dataset highlight three privacy risks. In future work we plan that the ontology will be used as part of a larger tool-chain, helping data publishers perform privacy assessments for datasets before data sharing.

9 Acknowledgments

This study was supported by the HBODEx project (64018-0558). The authors are participating in IEA EBC Annex 79 and were support by EUDP (64018-0558).

References

1. Ardakanian, O., Bhattacharya, A., Culler, D.: Non-intrusive techniques for establishing occupancy related energy savings in commercial buildings. In: BuildSys '16. pp. 21–30 (2016)
2. Arendt, K., Johansen, A., Jørgensen, B.N., Kjærgaard, M.B., Mattera, C.G., Sangogboye, F.C., Schwee, J.H., Veje, C.T.: Room-level occupant counts, airflow and CO2 data from an office building. In: Proceedings of the First Workshop on Data Acquisition To Analysis. pp. 13–14. DATA '18 (2018)
3. Arief-Ang, I.B., Hamilton, M., Salim, F.D.: A scalable room occupancy prediction with transferable time series decomposition of CO2 sensor data. ACM Trans. Sen. Netw. **14**(3-4), 21:1–21:28 (Nov 2018)
4. Arief-Ang, I.B., Salim, F.D., Hamilton, M.: Cd-hoc: Indoor human occupancy counting using carbon dioxide sensor data. arXiv preprint arXiv:1706.05286 (2017)
5. Balaji, B., Bhattacharya, A., Fierro, G., Gao, J., Gluck, J., Hong, D., Johansen, A., Koh, J., Ploennigs, J., Agarwal, Y., Berges, M., Culler, D., Gupta, R., Kjærgaard, M.B., Srivastava, M., Whitehouse, K.: Brick : Metadata schema for portable smart building applications. Applied Energy (2018)
6. Beckel, C., Sadamori, L., Staake, T., Santini, S.: Revealing household characteristics from smart meter data. Energy **78**, 397 – 410 (2014)
7. California State Legislature: California Consumer Privacy Act of 2018 (Jun 2018), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375
8. Carrara, W., Oudkerk, F., Van Steenbergen, E., Tinholt, D.: Open data goldbook for data managers and data holders (feb 2018)
9. Dwork, C.: Differential privacy. In: Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II. pp. 1–12. ICALP'06, Springer-Verlag, Berlin, Heidelberg (2006)
10. Dziedzic, J.W., Yan, D., Novakovic, V.: Real time measurement of dynamic metabolic factor (D-MET). In: Johansson, D., Bagge, H., Wahlström, Å. (eds.) Cold Climate HVAC 2018. pp. 677–688 (2019)
11. 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**, 1–88 (May 2016), https://eur-lex.europa.eu/eli/reg/2016/679/oj
12. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv. **42**(4), 14:1–14:53 (Jun 2010)

13. Kaur, M., Salim, F.D., Ren, Y., Chan, J., Tomko, M., Sanderson, M.: 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. pp. 130–139. BuildSys '18 (2018)
14. Khalil, N., Benhaddou, D., Gnawali, O., Subhlok, J.: Sonicdoor: Scaling person identification with ultrasonic sensors by novel modeling of shape, behavior and walking patterns. In: BuildSys '17. pp. 3:1–3:10 (2017)
15. Kjærgaard, M.B., Johansen, A., Sangogboye, F., Holmegaard, E.: Occure: An occupancy reasoning platform for occupancy-driven applications. In: CBSE 2016. pp. 39–48 (April 2016)
16. Kleiminger, W., Beckel, C., Santini, S.: Household occupancy monitoring using electricity meters. In: UbiComp '15. pp. 975–986 (2015)
17. Klyne, G., Carroll, J., McBride, B.: Rdf 1.1 concepts and abstract syntax. https://www.w3.org/TR/rdf11-concepts/ (Feb 2014), (Accessed on 10/18/2019)
18. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: L-diversity: privacy beyond k-anonymity. In: ICDE'06. pp. 24–24 (Apr 2006)
19. Nergiz, M.E., Atzori, M., Clifton, C.: Hiding the presence of individuals from shared databases. In: SIGMOD '07 (2007)
20. Office of the Australian Information Commissioner: Privacy Act 1988, Compilation No. 81. https://www.legislation.gov.au/Details/C2019C00241 (Aug 2019)
21. Pappachan, P., Degeling, M., Yus, R., Das, A., Bhagavatula, S., Melicher, W., Naeini, P.E., Zhang, S., Bauer, L., Kobsa, A., Mehrotra, S., Sadeh, N., Venkatasubramanian, N.: Towards privacy-aware smart buildings: Capturing, communicating, and enforcing privacy policies and preferences. In: ICDCSW. pp. 193–198 (2017)
22. European Parliament, Council of the European Union: Directive 2002/58/EC of the european parliament and of the council of 12 july 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (directive on privacy and electronic communications). Official Journal of the European Union **L201**, 37–47 (Jul 2002), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2002:201:TOC
23. Pfenninger, S., DeCarolis, J., Hirth, L., Quoilin, S., Staffell, I.: The importance of open data and software: Is energy research lagging behind? Energy Policy **101**, 211–215 (02 2017). https://doi.org/10.1016/j.enpol.2016.11.046
24. Prudhommeaux, E.: SPARQL query language for RDF. http://www. w3. org/TR/rdf-sparql-query/ (2008)
25. Rashid Asmaa, H., Mohd Yasin, N.: Privacy preserving data publishing: Review. International Journal of Physical Sciences **10**, 239–247 (Apr 2015)
26. Rocher, L., Hendrickx, J.M., de Montjoye, Y.A.: Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications **10**(1), 3069 (2019)
27. Sangogboye, F.C., Arendt, K., Singh, A., Veje, C.T., Kjærgaard, M.B., Jørgensen, B.N.: Performance comparison of occupancy count estimation and prediction with common versus dedicated sensors for building model predictive control. Building Simulation **10**(6), 829–843 (Dec 2017)
28. Schwee, J., Sangogboye, F., Kjærgaard, M.: Evaluating practical privacy attacks for building data anonymized by standard methods (Apr 2019), IoTSec '19
29. Sonta, A.J., Jain, R.K.: Inferring occupant ties: Automated inference of occupant network structure in commercial buildings. In: BuildSys '18. pp. 126–129 (2018)
30. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05), 557–570 (2002)