Ontology of active and passive environmental exposure

Csilla Vámos¹⁺, Simon Scheider¹, Tabea Sonnenschein⁰ and Roel Vermeulen⁰

¹ Human Geography and Spatial Planning, Utrecht University, Netherlands
E-mails: c.k.vamos@uu.nl, s.scheider@uu.nl
² Institute for Risk Assessment Sciences, Utrecht University, Netherlands
E-mails: t.s.sonnenschein@uu.nl, R.C.H.Vermeulen@uu.nl

Abstract. Exposure is a central concept of the health and behavioural sciences needed to study the influence of the environment on the health and behavior of people within a spatial context. While an increasing number of studies measure different forms of exposure, including the influence of air quality, noise, and crime, the influence of land cover on physical activity, or of the urban environment on food intake, we lack a common conceptual model of environmental exposure that captures its main structure across all this variety. Against the background of such a model, it becomes possible not only to systematically compare different methodological approaches, but also to better link and align the content of the vast amount of scientific publications on this topic in a systematic way. For example, an important methodical distinction is between studies that model exposure as an exclusive outcome of some activity versus ones where the environment acts as a direct independent cause (active vs. passive exposure). Here, we propose an ontology design pattern that can be used to define exposure and to model its variants. It is built around causal relations between concepts including persons, activities, concentrations, exposures, environments and health risks. We formally define environmental stressors and variants of exposure using Description Logic (DL), which allows automatic inference from the RDF-encoded content of a paper. Furthermore, concepts can be linked with data models and modelling methods used in a study. To test the pattern, we translated competency questions into SPARQL queries and ran them over RDF-encoded content. Results show how study characteristics can be classified and summarized in a manner which reflects important methodical differences.

Keywords: ontology, epidemiology, Python, RDF, health, GIS, computer science

1. Introduction

There is an increasing amount of work measuring some form of exposure to the environment to study its spatial effects on a person’s behaviour and health (cf. [1]). Yet the increasing amount and variety of approaches make it very time-consuming for researchers to find and compare results across articles relevant to some analytic goal. For example, a health-related study on walking behaviour might target the effects of outdoor air pollution while walking, or measure the effects of green space on walking behaviour, or the effects of such behaviour on physical health. Which goal precisely was pursued is hard to tell from a distance. While some authors have emphasized the opportunities of a corresponding "spatial turn" in the health sciences [2, 3], others, therefore, see the increasing need to synthesize such evidence and systematically structure underlying models with the help of information ontologies [4]. This may allow systematic comparisons of the effects of interventions on behaviour and health, and thus support evidence-based theory building [5].

*Corresponding author. E-mail: c.k.vamos@uu.nl.
Information ontologies provide a way to make the shared conceptualizations underlying a particular kind of information explicit [6]. Since such conceptualizations can differ greatly even between research on the same topic, understanding them is crucial for validating and comparing research results. Over the past couple of decades, ontologies, therefore, have become increasingly useful across medical and epidemiological sciences [7, 8]. To make conceptualizations explicit, ontologies make use of formal logic, which not only helps unambiguously define ideas (contributing to theory development) but also makes definitions machine-readable and thus helps automatically classify results (contributing to comparison and information retrieval). Together with methods for extracting and annotating content in published texts, this methodology can be used to link various resources underlying exposure studies. However, an ontology for breaking down and organizing different exposure concepts is currently lacking (cf. Sect 2).

Systematically distinguishing and aligning exposure measurements involves two major challenges: For one, there is an ontology design challenge [9, 10] to capture the central differences in the way exposure is modelled and used in scientific studies, such that we can answer corresponding questions [11, 12]. One important kind of question is about one’s level of control of exposure. It asks whether the health exposure studied is largely under the subjects’ control or not. The former we call active exposure and applies, for instance, to the exposure to unhealthy food, whereas the latter is called passive, for example, when being exposed to air pollution [13]. In the former case, buying or eating food is an activity that causes an exposure which can potentially be controlled by the involved person, while in the latter case, such control is not possible (furthermore, there are different types of passive exposure, which will be further explained in section 4.1.3). This distinction\(^1\) is relevant because it determines which model components are required. For passive exposure, tracking of people’s (mobility) behaviour and environmental stressor concentrations need to be modelled in detail, while for active exposure, behavioural choices of humans move into focus [13]. The distinction also has ethical and intervention/policy implications, because it determines to what extent a health impact due to exposure can be attributed to a person’s responsibility. However, to date, it remains unclear how this distinction can be precisely defined and operationalized. In addition, to capture the type of exposure and the tools and data sets used, we also need to identify the involved types of activities and subjects, their involved risk and the underlying environmental factors or “stressors” and how they were modelled. The second challenge relates to knowledge extraction, namely how data for such an ontology can be extracted, and how this can be scaled up across many article documents. Manually annotating articles with ontology concepts is a time-consuming process which does not scale. Luckily, recent developments in Natural Language Processing (NLP), such as the development of pre-trained deep neural networks for language parsing [14] have vastly increased the chances of automating the detection of exposure concepts within article texts[15].

Since this latter challenge requires first addressing the former, we concentrate in this paper on the first step of ontology design: Which concepts are needed to define exposure in epidemiological and health geography studies, such that relevant methodical differences like the types of exposure, environmental stressors and activities can be distinguished, including the underlying tools and data sources used for modelling it. In the following Sect.2, we discuss related work and requirements for such an ontology. In Sect. 3, we explain our design method, and in Sect. 4, we introduce the conceptual model, its (Web Ontology Language) OWL axiomatization and our reuse of existing vocabularies. Finally, in Sect. 5, we test and evaluate our ontology pattern over sample articles for these requirements.

2. Ontologies of exposure and competency questions

In this section, we review related work on exposure-related ontologies and formulate requirements for ontology design in terms of competency questions.

\(^1\)Sometimes also captured by the dichotomy voluntary vs. involuntary exposure. We prefer active/passive over voluntary/involuntary because the latter additionally implies an intention of the involved person, which we think is too restrictive.
2.1. Approaches to modelling health-related exposure and the environment

Information ontologies can be used to structure information in epidemiology and related fields [8]. Facts can be organized e.g. in terms of a so-called knowledge graph [16], which can be used to query, link or embed knowledge in various AI systems, e.g., for deep learning-based Natural Language Processing (NLP) and information retrieval [17]. However, conceptualizations, as well as terminology, can differ greatly not only between different fields but also within a single field, such as bio-medicine [18]. Designing large general-purpose domain ontologies, as was often done in the past, has therefore turned out to be difficult [19]. More recently, researchers have therefore turned to model aspects of a knowledge domain in terms of small, reusable design patterns for particular purposes [20], e.g., based on the types of questions they can answer [11]. Patterns can then be linked to form larger ontologies for specific purposes. Our ontology fits this smaller scale, with a focus to systematically compare methodological approaches with the aim to better link and align the content of the vast amount of scientific publications on exposure epidemiology. This ontology provides a good example of the power of designing ontologies for a particular purpose, but that can still be connected to other ontologies.

The concepts underlying environmental exposure may serve as a central pattern in this sense, linking domains such as epidemiology, environmental science, geography and behavioural sciences. Yet, researchers have modelled exposure from different angles in the past. In the following, we review ontologies and limitations that may arise, in the fields of biomedicine, healthy living, and epidemiology, as well as on particular exposure-related health factors, such as food, physical activities, as well as human behaviour. Finally, we discuss the only existing ontology that specifically focuses on exposure.

The Ontology for Biomedical Investigations (OBI ontology) [21] for biomedical investigations is an example of a large general domain ontology. In a multidisciplinary field posing challenges to terminology agreement, OBI suffers from corresponding problems. External ontologies reused in OBI are often subject to change with independent release policies, which can impact the scalability of changes to OBI [21]. For our purpose, the ontology is too general to address the specific problem of modelling exposure.

Various ontologies focus on medical health services, such as the one by [22, Bickmore et al]. The authors explore the possibility of using ontology to counsel patients on adopting a healthier lifestyle. Since the ontology is focusing on the cognitive requirements of human interactions, it is less suitable for exposure assessments. Another example is the medical ontology by Zeshan and Mohamad [23], which was designed to aid in making rapid, crucial decisions in healthcare. Zeshan and Mohamad’s ontology does not capture exposure concepts. Similarly, the ontology by Chen et al., [24], concerns the treatment and diagnosis of diabetes but does not include exposure as a concept.

[7, Pesquita et al] noted that many epidemiology-related ontologies have described concepts of specific subdisciplines such as the Disease ontology, Vaccine Ontology, and Symptom Ontology. In these ontologies, important epidemiological concepts are not yet covered, such as “exposure ratio” and “attack rate” [7]. The authors, therefore, created a general domain ontology called The Epidemiology Ontology (EPO) which covers some of these gaps [7]. The ontology also models exposure, but not in terms of a general environmental concept. Rather, it regards exposure as a process of transmission of infectious or other disease agents among persons².

Several ontologies focus on modelling the food environment. FoodOn is an ontology that covers basic raw food source ingredients, and process terms for packaging, cooking, and preservation. It also includes an upper-level product type scheme under which food products can be categorized. This ontology helps describe and organise food in detail and has been successful in standardizing database content for food-related agencies and health organizations [25]. The NAct ontology by [26, Tsatsou et al] focuses on connecting data about activities and nutrition. While many nutrition models already exist, NAct takes a holistic approach by combining and personalizing nutritional and physical activity recommendations to support healthy living. The authors adopt rules which connect each subject’s implicit and explicit nutritional and well-being goals with the situational condition of the subject, as well as with standardized European nutritional and well-being directives [26]. Both ontologies may be useful to model aspects of a food environment but they lack notions of exposure.

²EPO defines exposure as a BFO span:ProcessualEntity with the informal description “Proximity and/or contact with a source of a disease agent in such a manner that effective transmission of the agent or harmful effects of the agent may occur.”
ORBM+ [27] is an ontology that models human behaviour. The authors study how social relationships and personal factors contribute to macro-level behaviours, such as physical exercise. They developed the ontology using a knowledge-driven approach, followed by a data-driven validation and refinement approach. The key idea is that a representation of a concept will be learned by its own properties, the properties of its related concepts, and the representations of its sub-concepts [27]. This ontology is linked to a human behaviour deep learning prediction model to make the behaviour prediction explainable. By incorporating human behaviour determinants – self-motivation, implicit and explicit social influences, and environmental events, the model predicts the future activity levels of users more accurately than conventional methods [27].

[28, Mattingly et al] address the general conceptual challenges of exposure science with the ExO ontology. The authors note that while exposure-related terms are widely used in exposure science, definitions and descriptions are often inconsistent. The ontology is used to translate findings in various environmental disciplines, including epidemiology, for exposure and risk assessment and decision making and for improving public health [28]. The authors base their ontology on the gene ontology project, an ontology that describes the functions of gene products from all organisms [29]. ExO is structured hierarchically to allow the representation of data and concepts at varying levels of detail [28]. Mattingly et al suggest that the essence of exposure science is the study of the co-occurrence of a stressor and a receptor or a target. However, as we will explain below, reducing exposure to cases induced by stressors is too narrow, since not in all cases, stressors or targets that receive the impact of a stressor are available.

Also, ExO lacks formal definitions of exposure and related concepts that can be used to automate the classification of different types of exposures, such as passive and active exposure.

While all ontologies discussed above touch on some aspects of exposure, including the behavioural component, different kinds of environmental stressors, as well as more general medical terms or risks, it is still unclear how concepts fit into each other when determining and measuring exposure. Furthermore, it also seems that even existing exposure ontologies such as ExO are not general enough and thus fail to capture important variants of exposure, e.g., the difference between active and passive exposures or environmental factors that are not stressors but that beneficially affect people. The ontology which we propose in this paper exactly addresses this gap by taking the different components underlying exposure measurement into focus.

2.2. Competency questions about constituents and types of exposure measurement

As our discussion illustrates, ontologies relevant to exposure are ranging from understanding human behaviour and classifying physical activities and chemical substances, to the kinds of nutrition and their effect on people’s health. At first look, these cases are hard to align with each other in one model. Secondly, there are important differences between exposure measurements in terms of methods and data. Given this variety, the question is what an overarching model of exposure could look like which can be reused across all these cases to answer fundamental questions about methodology.

To capture such requirements, we formulate competency questions [12]. Competency questions should include those types of questions that an ontological model of exposure should be able to answer across all applications. We focus on understanding the conceptual model used in an article, and how it serves to link the used methods and data sources. Below, the rationale for each question is explained.

**Question 1. What kinds of exposure are modelled in this article?**

The variety of health-related exposures needs to be distinguished systematically and automatically. Identifying which exposure concept is used in a paper helps the reader determine if the article is of relevance. It also determines which environmental and individual aspects are relevant for modelling.

**Question 2. Which activities are involved in the exposure and who is exposed?**

Activities are causing exposure of the people involved in them. At the same time, different kinds of activities mediate exposure. For instance, walking to school may cause higher exposure to air pollution than driving to school for the same route. Children that need to walk 2km to school will have a higher exposure to physical activity than children that only need to walk 800 meters to school. Additionally, the health conditions of people involved in an activity can also influence how they react to exposure. For instance, children may be more susceptible to NO2 than adults. Thus, the demographic characteristics of persons and their activities both modify their health risks.
Question 3. What are subjects exposed to?

Whenever we are exposed to an environment, we are exposed to something specific. Depending on what this is, the exposure can be quantified in different ways, which determines the specific kind of exposure. Note, what a person is exposed to is not necessarily that person’s environment. For example, a person’s exposure to unhealthy food is not directly caused by the environment but is rather a consequence of an eating activity which is influenced by some (eating or buying) decisions in the environment.

Question 4. What is the health risk of exposure?

This question identifies the potential health risk an exposure may have for a subject. For instance, the risk of children that live near busy roads developing asthma. Note that exposure can either decrease or increase risk and thus may have positive or negative associations.

Question 5. Which environmental factors influence the exposure and from which data sets were they derived?

Environmental factors are aspects of the environment which either directly or indirectly influence the exposure of a person. Depending on whether the exposure promotes risk or not, the effect of the environment on health can be positive or negative. For instance, a negative environmental factor could be high temperatures, which are themselves caused by impervious surfaces, in what is called the heat island effect in a city [30]. An example of a positive environmental factor would be a park in a city increasing a citizen’s recreational activity, which in turn reduces the risk of obesity. Environmental factors are often derived from measurements of environmental phenomena (e.g. mean temperatures or object densities in a neighbourhood around the home). This means the analytic methods involve a GIS workflow which derives spatial and temporal measures from environmental layers. Therefore, if available, we are also interested in the workflows used for measuring these environmental factors, including the data sources.

Question 6. What are the environmental stressors?

An environmental stressor is an environmental factor that negatively influences the health risk of a person via her exposure. For example, high temperatures and impervious surfaces can be environmental stressors for elderly people in a city.

3. Methodology

In this section, we explain our approach to developing and evaluating the ontology pattern. Fig. 1 shows an overview of the development process. Ontology design methods [10] usually start with requirements and purposes. Following the idea of pattern development [6, 11], these requirements are captured by competency questions. Based on these, we conceptualized a preliminary pattern in terms of open slots [9] for classes and relations (OWL object properties) which capture the distinctions needed for answering the questions, and then formalized the pattern in OWL 2. As an empirical basis, we selected six articles that covered diverse kinds of exposure and risk. Articles were selected from literature databases such that they should cover varying epidemiological risk factors and exposure types. We selected six articles on exposure to food, air quality, crime, active transport, and physical activity (see Table 1). We made sure the articles that cover the same risk factor have different underlying exposure concepts. We also made sure to have a diversity of both active and passive exposure examples, including passive exposure examples of perceptual and physical nature. A short description of each article can be found in the Appendix.

We then filled the slots of the pattern with examples manually extracted from exposure articles. If needed, the ontology was extended with new concepts. As far as possible we inherited classes from existing ontologies. The ontology design was done iteratively in several rounds revising the ontology based on the content of the articles. We then fully encoded the content of each article into RDF, using classes and relations from the pattern. Using a mature

4. https://pubmed.ncbi.nlm.nih.gov. However, note that a systematic review was beyond the scope of this article.
5. https://www.w3.org/RDF/
Fig. 1. Steps in building and evaluating the ontology pattern. Numbers refer to corresponding sections in the article.

Table 1 The content of six articles was used for the development and evaluation of the ontology.

<table>
<thead>
<tr>
<th>Title of article</th>
<th>Main Authors</th>
<th>Year Published</th>
<th>Health Exposure</th>
<th>Health Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Associations of Area-Level Violent Crime Rates and Self-Reported Violent Crime Exposure with Adolescent Behavioral Health</td>
<td>Grinshteyn et al. [31]</td>
<td>2018</td>
<td>crime, violence</td>
<td>Adolescents, Behavioral health, Mental health</td>
</tr>
<tr>
<td>Constituents of household air pollution and risk of lung cancer among never-smoking women in Xuanwei and Fuyuan, China</td>
<td>Vermeulen et al. [32]</td>
<td>2019</td>
<td>(household) air pollution</td>
<td>lung cancer</td>
</tr>
<tr>
<td>Long-term exposure to air pollution and cardiorespiratory disease in the California teachers study cohort</td>
<td>Lipsett et al. [33]</td>
<td>2011</td>
<td>air pollution</td>
<td>cardiorespiratory diseases</td>
</tr>
<tr>
<td>Neighbourhood fast food exposure and consumption: The mediating role of neighbourhood social norms</td>
<td>Van Rongen et al. [34]</td>
<td>2020</td>
<td>fast food outlets, neighbourhood social norms</td>
<td>fast food consumption</td>
</tr>
<tr>
<td>The relationship between access and quality of urban green space with population physical activity</td>
<td>Hillsdon et al. [35]</td>
<td>2006</td>
<td>urban green space</td>
<td>physical activity levels</td>
</tr>
<tr>
<td>Natural and built environmental exposures on children’s active school travel: A Dutch global positioning system-based cross-sectional study</td>
<td>Helbich et al. [36]</td>
<td>2016</td>
<td>natural and built environment, travel mode</td>
<td>activity level of children</td>
</tr>
</tbody>
</table>

version of the ontology, we then automatically enriched the RDF-encoded article contents by running OWL-RL and RDFS inference over the data. This step adds automatic class instantiations to the article content based on the formal definitions specified in the ontology pattern, and in this way allows us to classify article content based on logical reasoning (e.g. the fact that a certain exposure is of a certain type).

To evaluate the pattern, we translated the competency questions into SPARQL queries and finally ran all queries over the enriched article contents to analyse the content and to automatically classify and compare the articles against each other. We also compared the result against our expectations from reading the articles. This tests two things: first, whether the pattern is general enough to cover the diversity of exposure methods and specific enough to distinguish important methodical differences. And second, to what extent the pattern can be used for retrieval of methodological content? We discuss the results in Sect 5.

6https://www.w3.org/TR/owl2-profiles/#Reasoning_in_OWL_2_RL_and_RDF_Graphs_using_Rules
7https://www.w3.org/TR/rdf11-mt/#rdfs-interpretations
8https://www.w3.org/TR/rdf-sparql-query/
4. Ontology design

This section describes our ontology design, motivating concepts and the types of relations used to build it with the aid of description logic axioms. Description Logic (DL) is implemented in the W3C standards OWL and RDF. Many fragments of this logic are decidable and thus allow not only defining classes and relations between classes, but also the automatic inference of class subsumption (whether classes are subclasses of each other), and class instantiations (whether e.g. data samples can be classified accordingly). The ontology pattern was tested for consistency/coherency using the HermiT reasoner.

Ontologies are often divided into upper/top-level and domain ontologies, as well as lightweight and heavyweight ones. Lightweight ontologies are mere taxonomies. Upper ontologies axiomatize general categories that can be reused across many knowledge domains. An example of an upper-level ontology is the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) ontology. DOLCE embraces a pluralist, cognitive perspective rather than targeting a unique universal ontology for knowledge representation.

Ontologies may also be built off one another, similar to design patterns in software engineering. Our ontology pattern can be used across many domains concerned with exposure, and it goes beyond a mere taxonomy because it defines classes based on causal structures. We aligned our classes with the top-level ontology DOLCE+DnS Ultralite ontology (DUL), as it includes basic ontological distinctions relevant for modelling environmental agency (discussed below). In addition, we inherited from the EPO:exposure class. We also reused a recently published ontology on quantities (AMMO and GeoAMMO) to describe quantifiable measures of exposure. Our pattern exposureBasis (expB) is available online as well as on github together with all resources.

4.1. Basic model of active and passive exposure

We start with an informal description before introducing formal definitions. We first discuss the role of causal relations in exposure measurement, before we introduce conceptual slots (classes) for the phenomena involved, and how they are related to each other. Afterwards, we introduce exposure types that can be defined as classes.

4.1.1. Causal relations and measure-able phenomena

From an analytical perspective, exposure is an important cause for risks or benefits regarding one’s health. For example, exposure to an environment can cause a particular behaviour (e.g. when we are triggered by a nearby park to go running), which can be an indirect cause of a health risk or health benefit (e.g. exercise and time outdoors being beneficial). Furthermore, it can also be a direct cause of health risks, e.g. when we are running near a busy road. Finally, the environment can be modified by behaviour (e.g. when we decide to take a car instead of walk).

Thus in environmental exposure, the environment can occur both as cause and effect in various causal chains. In general, causal relations link measurable phenomena in a way that allows us to model and predict them beyond spurious correlations. From causal theory, we know that measurable phenomena might not correlate even though there is a causal link between them, and vice versa. This is especially relevant for the environmental and health sciences. For example, whether the environment causes risks might be hidden by confounding effects (causal forks), such as residential self-selection. The distinction between causal and non-causal relations cannot be made without background assumptions. Making such assumptions explicit results in a causal diagram, where causal relations appear as directed arrows between measurable variables. In essence, such a diagram is a conceptual model which can be formalised in an ontology. For this reason, we use a generalized causal relation as a basic primitive DL role for connecting exposure phenomena.

---

9 For an introduction to the DL syntax, see cf. [37].
10 http://www.hermiT-reasoner.com/
11 http://ontologydesignpatterns.org/wiki/Ontology:DOLCE+DnS_Ultralite
12 geographicknowledge.de/vocab/Ammo.ttl
13 geographicknowledge.de/vocab/GeoAmmo.ttl
14 http://geographicknowledge.de/vocab/exposureBasis.ttl
15 https://github.com/simonscheider/exposureStudy
Which measurable exposure phenomena should be linked by causal relations? DOLCE and other top-level ontologies distinguish events (phenomena measured in time) from objects (phenomena measured in space), and causal relations typically exist only between consecutive events [40]. Other types of relations, e.g. participation, are used to link events and objects. However, the practice of causal analysis seems to be tolerant allowing causal links between all categories (e.g. an object like a park can cause an event like a run). We think this practice has also important theoretical implications because it highlights the role of particular causal chains for the conceptualization of exposure. More specific ontological relations can still inherit from general causality if needed. For example, we might specify that the person who decides to walk not only is a cause of the walking event but is also participating in this event. In the following, we leave the type of causal relation open to sub-patterns of the ontology. For example, there is a causal relation between both food intake and health risk, and noise and health risk. However, they are based on very different physical processes that might be specified further in sub-ontologies.

4.1.2. Exposure, Activity, Environment, Risk and Dose

What exactly is exposure? In the following, we base our explanations on the notions of measurement control as introduced by Sinton [48], as well as on standard definitions in epidemiology. Without being too specific, we can say that exposure is a measurement of some amount which is controlled by (and adds up) over time. Exposure refers to the amount of a particular environmental phenomenon that reaches an organism, expressed in terms of physical state, concentration, duration, and frequency. Exposure is a critical determinant of the probability of human harm (i.e. risk). If you are exposed to some phenomenon for some time, and then again for another time, the total amount of exposure will increase by the amount of exposure in this additional time interval. Exposure therefore can be defined as an temporally extensive amount, i.e. an amount controlled by an amount of time [42]. This amount of time is, in turn, controlled by some activity of the person who is exposed. For example, the amount of exposure to NO2 and the amount of physical activity are both controlled by the time interval of a person biking along a road with traffic. The longer a person bikes, the more exposed the person is to both.

An activity happens in time and involves a person. We hold that exposure is always measured relative to some activity; e.g. it is always based on the duration of the activity and can be measured relative to the location of the person involved in the activity. Yet, how the activity influences exposure is different for different types of exposure: in the case of food intake, the amount and the quality of food are important. In the case of noise, the duration and the location of the person involved are relevant. As in the example above, activities are caused by persons. This could be anything from simply living in a certain place, to biking, or to buying food. Activities can be stationary or involve movement. If persons have control over an activity they have the freedom to act (for example, you choose to smoke or not). Sometimes there are many alternatives to choose from (for example, for your commute to work, you can choose to bike, take public transport, or walk), and the decision to act is dependent on the environment. However, sometimes people do not have control over performing an activity. It is possible that the environment constrains people’s activity options, sometimes up to the degree that there is no choice and the activity becomes involuntary. In that case the person does not cause the activity but the environment or biological need causes it (e.g. person falling asleep because of exhaustion, or a person shivering because it is cold outside). In the following we assume activities are not necessarily voluntary.

An environment consists of characteristics of a neighbourhood near the location of a person. This could involve tangible phenomena of the landscape (road intersections, coal mines, fast food outlets) or intangible ones (NO2, farm odour), or even flat phenomena like a culture or an administrative boundary. What is considered an environment is therefore not only dependent on the spatial scale of a person, but it also depends on the person’s activity. For example, the environment for shopping is constrained by the accessibility of shops. Thus activities become constitutive of environments. It is this connection between the activity, environment, and exposure which leads to risks or benefits.

Dose is an amount of something accumulated in the person’s body due to exposure. For example, it can be the amount of a passive stressor (e.g. NO2, noise) that enters a person’s body dependent on the concentration or intensity in the environment and the physiological properties of the person.

\(^{16}\) Cf. https://www.endocrinescience.org/glossary/exposure/. We generalized “substance” in this standard definition to “phenomenon”, since some exposure types are immaterial.
Risk is a person’s probability of participating in an event that negatively influences the person’s health within a specified period of time\(^7\). For instance, a risk could be a heart attack, disruptive behaviour, or obesity. The degree to which a risk influences one’s health or mortality varies. However, note that exposure might also decrease risk. Health risks are often mediated by doses.

### 4.1.3. Active and passive exposure in a nutshell

While we suggest it is always these same basic components that constitute the parts of any exposure assessment, their causal configuration differs from one case to another. The argument we want to make in this article is that this causal structure precisely distinguishes the different cases of exposure from a methodological viewpoint. For example, for modelling the exposure to food intake we have to consider the environment (e.g. fast food outlets), some activity (e.g. buying, eating), and the characteristics of a person involved (e.g. age), as well as the risk involved (e.g. obesity). Analogously, to model the exposure to noise, we likewise have to take into account an environment (noise level and noise sources), some activity (commuting to work) in which some person (school child) is involved, as well as some risk (e.g. mental health). Thus, to be general enough to model various cases of exposure, our pattern needs to allow the modelling of the configuration of the causal relations between these components. For example, in the case of noise exposure, the exposure is caused by the environment. So there is a direct causal influence of the intensity of noise in the environment on the amount of exposure to noise, which then influences the amount of risk. So we have a chain: Environment → Exposure → Risk. Yet the activity likewise influences exposure, in the sense that it determines the spatial context of the person being exposed. Thus, both the activity directly influences the exposure to noise as well as the noise level: Activity → Exposure → Risk.

This is different in the case of food exposure, which is an exposure to an activity. In the latter case, the environment (e.g. the number of fast food restaurants) still plays an important role, but not as a direct cause of the exposure to food intake. This is reflected by the fact that no matter how many fast-food restaurants are around you, you are not forced to eat there. There is always an intermediary activity (and thus an implicit decision of eating or buying) involved between the environment and the exposure and the risk. For this reason, it is not in itself risky to drive by a McDonald’s restaurant, whereas it is risky to drive by a polluted area (such as a factory). Thus, for the food case, we instead have a chain: Environment → Activity → Exposure → Risk.

We call the second causal configuration active exposure, where the exposure and the risk are entirely controlled by a person, even though the person’s activity might be influenced by the environment. Note that this distinction has important implications for (1) the modelling of exposure (which components need to be modelled, in which order), but also in terms of (2) ethics: while fast food restaurants can be avoided, no one can avoid noise around an airport when driving by. The first causal configuration is called passive exposure. Though an activity is always involved, there is also a component of the risk which is entirely independent of a person’s activity (and thus beyond that person’s control). Depending on how this component affects a person’s body, exposure can be further distinguished into physiological exposure, and perceptual exposure. Physiological exposures include exposures that physically enter or affect the body (e.g. air pollution, sunlight). Perceptual exposures are exposures that involve perception (e.g. the perception of crimes and its effect on the feeling of safety).

We made the difference between active and passive exposures explicit in terms of the elements of our basic exposure ontology pattern, as depicted in Fig. 2. As can be seen, active and passive exposure is a matter of different causal links between classes. All classes can be linked by causal relations. We believe that all classes are relevant at least as background assumptions in a specific model. Though such assumptions may not explicitly be modelled with data, they are still playing a role in the exposure model. Yet, how do we make these distinctions formally explicit, such that a computer can compute with them?

### 4.1.4. DL-Axiomatization of exposure concepts

We first introduce base classes for the different open slots in our causal model of exposure (standing for the ellipses of Fig. 2), including exposure, environment, activity, person, dose and risk. These classes are all mutually exclusive, meaning that something cannot be of more than one of these classes at the same time (e.g. not a person and an activity):

\(^7\)Risk in epidemiology is commonly used more loosely to talk about probabilities of events more generally. We stick to a more restrictive interpretation which we think is more informative.
Fig. 2. The basic elements of the ontology. Arrows are causal relations, ellipses are classes of concepts. A conceptual model of exposure can use one or more of the depicted elements.

Axiom 1. Base classes are mutually disjoint

\[
\text{(Exposure} \cap \text{EnvironmentalFactor}) \sqcup (\text{Exposure} \cap \text{Activity}) \sqcup (\text{Exposure} \cap \text{Person}) \sqcup (\text{Exposure} \cap \text{Dose}) \sqcup \\
(\text{Exposure} \cap \text{Risk}) \sqcup (\text{EnvironmentalFactor} \cap \text{Activity}) \sqcup (\text{EnvironmentalFactor} \cap \text{Person}) \sqcup \\
(\text{EnvironmentalFactor} \cap \text{Risk}) \sqcup (\text{EnvironmentalFactor} \cap \text{Dose}) \sqcup (\text{Activity} \cap \text{Person}) \sqcup \\
(\text{Activity} \cap \text{Risk}) \sqcup (\text{Activity} \cap \text{Dose}) \sqsubseteq \perp
\]

Note that the phenomena that fall under these classes have measurable qualities that are not identical to the phenomena themselves. We distinguish different kinds of phenomena and their qualities using DOLCE+DnS Ultra-lite (DUL). Objects are phenomena whose qualities are controlled by time moments (DUL:Object). For example, persons (DUL:Person) as well as environments can change their qualities in time. We model environments as a DUL:PhysicalPlace. Activities are (DUL:Events), i.e., entities whose qualities are not controlled by time moments, but which have some fixed temporal extent. The following axioms specify causal relations (arrows) between the measured qualities of exposure concepts:

Axiom 2. Causal roles

\[
\text{causes} \equiv \text{causedBy} \quad \text{Disjoint}(\text{hinders, promotes})
\]

\[
\text{hinders} \equiv \text{isHinderedBy} \quad \text{hinders} \sqsubseteq \text{causes}
\]

\[
\text{promotes} \equiv \text{isPromotedBy} \quad \text{promotes} \sqsubseteq \text{causes}
\]

We consider a single causal relation causes which is the inverse of causedBy, denoting whether some quality of some phenomenon is causally influenced by some quality of another phenomenon. For example, both the environmental concentration of NO2 (a quality of some environment) and the duration of the cycling activity of a person (a quality of some activity) cause an exposure to NO2 (an accumulation amount). This causes a measure of the dose of NO2 in this person’s body. We only distinguish two sub-relations: hinders, which means a measured quality influences the other in a negative direction (the more, the less) or not (promotes).

The classes Exposure and Dose correspond to a particular kind of amount (AMMO:Amount), namely an amount accumulated over (and thus controlled by) some time interval (GeoAMMO:AccumulationAmount). More specifi-

---

An accumulation amount is measured by an accumulation measurement function. The latter is controlled by amounts of time [42].
cally, an exposure corresponds to a person’s accumulated amount of exposure to something over some time interval during an activity in which the person is involved. The time interval can be the extent of the activity or any part of it. For example, residents are exposed to local air quality at any time during which they reside in the same place. In this case, the air quality at the place is measured by concentration, the persons are residents and the activity is living somewhere. A dose is an amount of substance left in a person’s body as a consequence of its exposure. For example, this could be the amount of PM10 in your lungs. An exposure magnitude might be measured as a temporal integral of intensities, e.g. as a sum of NO2 concentration values over some time interval. Strictly speaking, the measured magnitudes (e.g. in grams) are not identical with the amount (e.g. the amount of NO2 in the body) [42]. Yet, in our design pattern, we do not further distinguish this to keep the pattern as simple as possible.

We formalize this causal structure by requiring that exposures always depend on persons via some activity in which they are involved during the exposure19. For example, a person’s exposure to NO2 is caused by that person’s biking. This requires exposure to be always caused by exactly one activity which is caused by exactly one person:

**Axiom 3. Exposures are caused by activities, and activities are caused by persons**

\[
\text{Exposure} \subseteq ((\exists \text{causedBy}\text{Activity}) \land (\leq 1) \text{causedBy}\text{Activity})
\]

\[
\text{Activity} \subseteq ((\exists \text{causedBy}\text{Person}) \land (\leq 1) \text{causedBy}\text{Person})
\]

This makes sure that for every exposure there is a unique person who is exposed, as well as a unique activity. Next, we specify the effects of exposure on this person in terms of its dose and risk. We call an exposure or dose health-relevant if it causes some risk for this person. For example, exposure to fast food increases the risk of obesity. Note some exposures are not health relevant because no risk is involved. For example, a sign may have caused me to stop at a traffic light. Furthermore, we call the activities causing these exposures also health-relevant. We define this in terms of DL role restrictions.

**Definition 1. Health impacts**

\[
\text{HealthRelevantDose} \equiv (\text{Dose} \land \exists \text{causes}\text{Risk})
\]

\[
\text{HealthRelevantExposure} \equiv ((\text{Exposure} \land \exists \text{causes}\text{Risk}) \lor (\text{Exposure} \land \exists \text{causes}\text{HealthRelevantDose}))
\]

\[
\text{HealthRelevantActivity} \equiv (\text{Activity} \land \exists \text{causes}\text{HealthRelevantExposure})
\]

\[
\text{RiskPromotingDose} \equiv (\text{Dose} \land \exists \text{promotes}\text{Risk})
\]

\[
\text{RiskPromotingExposure} \equiv ((\text{Exposure} \land \exists \text{promotes}\text{Risk}) \lor ((\text{Exposure} \land \exists \text{promotes}\text{RiskPromotingDose}))
\]

\[
\text{RiskPreventingExposure} \equiv ((\text{Exposure} \land \exists \text{hinders}\text{Risk}) \lor (\text{Exposure} \land \exists \text{hinders}\text{RiskPromotingDose})
\]

If we know such exposures (or doses) promote risk rather than hindering it, meaning there is a promoting chain of causes from activity to risk, then we speak of a risk promoting exposure (dose).

The term environmental stressor has been defined in various ways by different researchers. Most of these definitions involve both an environmental factor and some (negative) response for the exposed person. For example, Killen et al., [50] describes a stressor as “any intrinsic or extrinsic factor that challenges individuals and obliges them to adjust behavior”. [51] defines environmental stress as “as the emotional, cognitive and behavioral responses to an environmental stimulus (or stressor)”. Thus “whether stress occurs is dependent on individual and contextual factors.” [52] shows how environmental stressors can be further categorized according to the degree of actionability over the environmental stressor (directly or indirectly), predictability of the stressor, and how salient or identifiable the stressors are.

We define stressors simply based on the causal relation between environmental factors and health risks of the person exposed. Stressors (environmental stressors) are environmental factors which promote some exposure that

---

19A more complete formalization of our exposure definition above would require modelling amount domains explicitly in the pattern. We have refrained from doing so to keep the pattern simple.
promotes some risk. Note that stressors therefore are not necessarily involved: For example, in the case of exposure to fast food, there is no environmental stressor involved, because the environment does not directly cause the risky exposure. Furthermore, there are also environmental factors that cause exposures which hinder risk and thus promote health, e.g., exposure to green space. Finally, note that our definition leaves room for all the stressor related concepts cited above, including controllability, cognitive and physiological responses. These can be accounted for by distinguishing corresponding relations between actions and the kinds of exposure involved (cf. the distinctions defined below).

**Definition 2. Environmental stressors**

\[
\text{Stressor} \equiv (\text{EnvironmentalFactor} \sqcap \exists \text{promotes.RiskPromotingExposure})
\]

Finally, we can define the difference between active and passive exposures based on distinguishing their causes in terms of involved activities, and thus in terms of personal responsibility. We first introduce a class *Active*, which is defined as something that is either itself an activity or caused by some activity:

**Definition 3. Active**

\[
\text{Active} \equiv (\text{Activity} \sqcup \exists \text{causedBy.Activity})
\]

Note that this class includes, besides activities, also *active environmental factors*, whenever the latter is caused by some activity. For example, when we burn coal in an oven without a vent, we cause air pollution in our homes. Now, we call an exposure *active* if has *only* active causes. We call an exposure passive if it is caused by some environmental factor:

**Definition 4. Active and passive exposures**

\[
\text{ActiveExposure} \equiv (\text{Exposure} \sqcap \forall \text{causedBy.Active})
\]

\[
\text{PassiveExposure} \equiv (\text{Exposure} \sqcap \exists \text{causedBy.EnvironmentalFactor})
\]

This definition builds on the following logical reasons: If the exposure is caused only by some activity, and thus by the causes of that activity (e.g. a person’s decision to act) (by Axiom 3), then we know there is no independent influence of the environment on the exposure, and thus the responsibility of exposure lies entirely within the hands of the person who controls the activity. Our definitions partially distinguish between these different models as illustrated in Figure 3. However, to fully implement this idea of active exposures in our model, we need to assure that the exposure is not caused by something which is not active. This requires knowing whether something is not the case (logical negation \(\neg\)), which requires the logical closure of our knowledge base (cf. [53]). Since DL has an open-world assumption (what we don’t know is not automatically false), this reasoning goes beyond standard DL reasoning. To account for this, we *locally closed our world* of causes to be able to make this inference within DL:

**Inference rule 1. Local closure of causedBy Activity. If something is caused only by activities in the graph g, then we add an all-constraint:**

```python
def locallyCloseWorld(g, property=expB:causedBy, all = expB:Active):
    for s in g.subjects(property, None):
        allconstraint = False
        objects = g.objects(s, property)
        for o in objects:
            if (o, rdf.type, all) in g:
                allconstraint = True
            else:
                allconstraint = False
                break
        if allconstraint:
            g.add(s, rdf.type, \forall property.all)
```
(a) Possible model of active exposure. Note that our definition does not exclude environmental influence, but restricts it to be at most an active cause of exposure.

(b) Possible model of passive exposure. Though termed "passive", some activities are always implied.

Fig. 3. Possible models of active and passive exposure allowed by our definitions.

Note that our two definitions for passive and active exposure are not mutually exclusive, namely, in case the environmental factor is caused by some activity (e.g. burning coal). To make them mutually exclusive, we would need to request that passive exposure factors are never active, which requires another local closure of a similar kind. In this paper, we decided to leave this stricter definition out, because the more loose definition also illustrates that sometimes exposure can be considered both active and passive. In addition, we added subclasses for perceptual and physiological exposure which capture differences in the way exposure is caused by its factors. The exposure class can be seen as a reified n-ary relation between the causes constituting the exposure. Thus, these exposure subclasses also capture specific ways in which environmental factors are related, e.g., via perception or via physiological contact. Note that we do not restrict this to passive exposures, since in principle, active exposures could be caused by perceptual or physiological causes which are themselves controlled by activities (such as burning coal). These exposure subclasses are likewise not mutually exclusive:

**Axiom 4.** Kinds of exposures

\[
\text{Perceptual} \subseteq \text{Exposure} \\
\text{Physiological} \subseteq \text{Exposure}
\]

We do not further specify environmental factors using an ontology because such phenomena can be very diverse and thus should be modelled in a separate effort.

4.2. Modeling data generation

The previous section introduced an ontology that can be used to reason over the different concepts that are needed to understand how exposure is modelled in an article. An important aspect of this question concerns how concepts are represented in terms of data.

In general, concepts may either stay implicit in the actual analysis or else may explicitly be represented by data. Certain factors involved in the exposure process are often part of the background assumptions without any explicit modelling. For example, many studies neither model the persons involved in exposure explicitly nor the
actual exposure event, while others leave the environmental factors implicit. Yet, still, these concepts are important
to understand the author’s intentions and methodological approach. To investigate the extent of explicit/implicit
modelling within an article, we indicate whether a concept has a data representation or not and if yes, from which
data sources they might have been derived if this is known.

For this purpose, we use the Data Catalog Vocabulary (DCAT)\(^{20}\), Version 2, which is used to describe datasets
and their distributions (via different URLs). To keep things simple, we label something as both an instance of a
concept and of data set (\textit{dcat:Dataset}), meaning that the respective data set is instantiating the concept. For example,
there might be a data set of temperature measurements which is at the same time an environmental factor. Using the
property \textit{dcat:distribution}, we link the dataset to a particular distribution source, e.g., some URI from a public data
catalogue.

Second, since data sources are often not used directly but need to be \textit{transformed} to capture information about
the intended concepts, we model such transformations by linking dataset nodes using the provenance ontology
(\textit{PROV})\(^{21}\), using the property \textit{prov:wasDerivedFrom} (in case the derivation method is unknown), or else by triples
of the following pattern:

\[
\text{_:dataOutput prov:wasGeneratedBy :a1 .}
\]
\[
_:a1 prov:wasAssociatedWith <https://cran.r-project.org/web/packages/kdensity/>.}
\]
\[
_:a1 prov:used :dataInput.
\]

, where \textit{:a1} denotes the application of some R tool to derive \textit{:dataOutput} from \textit{:dataInput}.

4.3. Encoding of example articles

All six articles were encoded to test the ontology. For this purpose, we first identified all article content/text
snippets which denote instances of some class in our ontology. In this study, we did this thoroughly by reading the
articles and manually identifying the text phrases which corresponded to classes in ExpB or the dataset/provenance
ontologies. We then generated a blank node for each detected text phrase that stands for an instance of a class (e.g.
\textit{expB:EnvironmentalFactor}) and saved the corresponding phrases into RDFS comments describing this blank node:

\[
\text{_:proportionofcyclingpathlengths a expB:EnvironmentalFactor, dcat:Dataset;}
\]
\[
rdfs:comment "proportion of cycling path lengths".
\]
\[
_:proportionofcyclingpathlengths prov:wasDerivedFrom :cyclingstreets.
\]

In the future, this work may be automatized using a larger annotated corpus of articles and state-of-the-art deep
learning-based NLP methods, similar to \cite{54}. A particular challenge is that the concepts that play a role in the
exposure assessment are sometimes left implicit by the authors. Furthermore, we added causal and other links
between extracted instances whenever the authors either gave support for such a link (e.g. if they found a correlation)
or when they mentioned or assumed such links in their overall approach. Both practices require implicit knowledge
and therefore currently still pose a challenge for state-of-the-art NLP methods \cite{54}. Since our article rather focuses
on the modelling aspect, we did not use state-of-the-art text annotation techniques for finding text snippets \cite{55}.

5. Evaluation: comparing conceptualizations and methods for measuring exposure

To evaluate the pattern, we tested to what extent the competency questions can be answered automatically in a
way that corresponds to our understanding of each article’s method.

\(^{20}\)https://www.w3.org/TR/vocab-dcat-2/
\(^{21}\)http://www.w3.org/ns/prov#
5.1. Translating competency questions into SPARQL queries

SPARQL\(^{22}\), the query language for RDF, is used here to automatically retrieve answers for competency questions. In the following, we go through each question and discuss its translation to SPARQL:

**Query 1.** 'What kind of exposures are modelled in this paper?'

```sql
SELECT DISTINCT ?c ?y
WHERE {
  ?x a expB:Exposure.
  ?x rdfs:comment ?c
  OPTIONAL {?y a ?y.
    FILTER(?y not in (expB:Exposure, dcat:Dataset)).
    FILTER(!isBlank(?y))}
}
```

Here we query for exposures (\(?x\)) and retrieve the other classes (\(?y\)) they are instances of (other than expB:Exposure and not dcat:Dataset, constrained by FILTER), in case they exist (OPTIONAL statement).

**Query 2.** 'Which activities are involved in the exposure and who is exposed?'

```sql
SELECT DISTINCT ?yc ?zc
WHERE {
  ?x a expB:Exposure.
  ?x expB:causedBy ?y. ?y a expB:Activity.
  OPTIONAL {?y expB:causedBy ?z. ?z a expB:Person.
    ?z rdfs:comment ?zc.}
}
```

In this query, we search for activities that cause some exposure, and optionally for persons that performed (caused) such an activity. Since the ontology does not involve any activity/person types, we just retrieve the text descriptions (rdfs:comment) about these activities or persons.

**Query 3.** 'What are subjects exposed to?'

```sql
SELECT DISTINCT ?yc
WHERE {
  ?x a expB:Exposure.
  FILTER NOT EXISTS {?x a expB:ActiveExposure. ?y a expB:EnvironmentalFactor.}
  FILTER NOT EXISTS {?x a expB:PassiveExposure. ?y a expB:Activity.}
}
```

In this query, we search for all phenomena that cause exposure. Yet, the focus on what we are exposed to changes with the type of exposure. In the case of passive exposure, we focus on environmental factors. This is because if someone is passively exposed to air pollution, e.g., we are not interested in his or her activity performed when being exposed. Conversely, for active exposure, we are mainly interested in the activity that is performed, such as running. This focus is encoded in FILTER NOT EXISTS statements, and of course, it could be removed if needed.

**Query 4.** 'What is their risk of exposure?'

```sql
SELECT DISTINCT ?yc
WHERE {
  ?x a expB:Exposure.
  ?x rdfs:comment ?c.
}
```

\(^{22}\)https://www.w3.org/TR/sparql11-query/
In this query, we retrieve risks caused by exposures, potentially via some causal chain (+). This is because the exposure may cause risks directly or indirectly via doses first. We want to keep this possibility open.

**Query 5.** ‘Which environmental factors influence the exposure and from which datasets were they derived?’

```sparql
SELECT DISTINCT ?yc ?zc ?d
WHERE {
  ?x a expB:Exposure.
  ?x rdfs:comment ?xc.
  ?y prov:wasDerivedFrom+ ?z. ?z a dcat:Dataset; rdfs:comment ?zc.
  FILTER NOT EXISTS {?z prov:wasDerivedFrom ?u}
  OPTIONAL{?z dcat:distribution ?d}
}
```

In this query, we search for environmental factors that (directly or indirectly) cause exposure. The causal chain (+) is needed since, in the case of active exposures, the environment is a direct cause of the activity, but only an indirect cause of the exposure, via the activity. Furthermore, we are also interested in the data sources of these environmental factors, which could have been generated by zero or more (+) steps of derivation via the provenance ontology `prov:wasDerivedFrom`. We want to focus on the sources of data, not intermediary datasets (FILTER NOT EXISTS) and possibly (OPTIONAL) retrieve a web link to where the data is available (`dcat:distribution`).

**Query 6.** ‘What are the environmental stressors?’

```sparql
SELECT DISTINCT ?xc
WHERE {
  ?x a expB:EnvironmentalFactor; rdfs:comment ?xc.
  ?y a expB:RiskPromotingExposure; expB:causedBy ?x .
}
```

In this query, we are looking for environmental factors that cause some risk-promoting exposure (cf. Definition 1), i.e., an exposure that causes a risk level to increase with the amount of exposure. This is what we call an environmental stressor.

5.2. Running inferences and queries

We loaded RDF files for each paper together with our ontology into separate RDF graphs in RDFLib. We then used a brute force implementation of the OWL 2 RL and RDFS inference schemes to expand each graph with all possible triples that logically follow from our ontology and the linked data encoding of a paper’s content. After this inference step, we applied locally closed world inferences to all unique `causedBy:Active` triples (as explained in Inference rule 1) using our script. Since the latter adds new OWL facts which serve as a start for further inferences, we needed to run the former inference steps again. Since the standard inference is conservative regarding `causedBy` triples, no further inference is possible. Afterwards, we fired all SPARQL queries over all graphs and summarized the answers.

5.3. Results

In this section, we discuss the potential of our model for filtering and classifying exposure-related concepts, data and methods across studies. For this reason, we compare results across the six studies for each query individually. Retrieved answers to queries are shown in Tables 2 and 3.
Query 1 As you can see in Table 2, the amount of answers in each study for this query already tells us something about the focus of a study. For example, [36], [34] and [35] only study a single exposure, whereas [33], [31] and [32] study multiple exposures. For example, [33] focus on types of air quality exposures and [31] on different variants of crime exposures. All exposures are health relevant. Furthermore, we can see differences in how these exposures are automatically classified using inference. According to our model, [35], [34], [36] and [32] all study some form of active exposure. According to Def. 4, this means that exposures have exclusively active causes (so either are activities or are caused by activities). [36] and [35] e.g. focus on exposure to physical activity (walking or biking or motorized transport), while [34] focuses on an individual’s exposure to poor diet and fast food. Note that while these studies also take exposure to environmental factors into focus, the latter are not direct causes of exposure. Furthermore, the poor diet exposure in the study of [34] is correctly classified as a risk-promoting exposure, whereas the other two kinds of exposures are correctly recognized as risk preventing instead. [32] is an interesting case of active risk-promoting exposure. Though air quality plays an important role in this process, the exposure is still classified as active, simply because burning coal is an activity causing air quality, and so the causal chain of exposure is entirely rooted in the underlying household decisions of the women. [31] is another interesting border case, because exposure to crime may be seen as an active exposure due to crime being an activity, yet it is classified as passive by our model. The reason is that [31] does not take into account crime as an activity, including the people committing the crime, but rather models crime as a (static) aspect of the environment. This way of modelling crime resembles the way any other environmental factor is modelled.

Query 2 asks about the specific activity that causes health-relevant exposure and who is involved in that activity (see Table 2). In the case of [31], this activity is not committing a crime, but living in a neighbourhood with crime, as experienced by children. Living is also the prime activity considered in the study of [33] about air pollution, yet in this case, focusing on female teachers. Children’s transport to school is the focus for [36], whereas [35] focuses on the physical activity of adults in Norwich. Interestingly, the activity causing the exposure in [34] is not the food buying behaviour (though this could be done when studying exposure to poor diet), but it is instead eating at fast food outlets. Note that this distinction is crucial to understand whether studies about food are comparable to not.

In [32], our model makes clear that the cause of the smoky coal exposure is indoor fuel use by never smoking women. This shows the study intends to measure a health effect that can be exclusively attributed to the household environment, instead of smoking behaviour.

As shown in Table 2, most studies only take a single kind of activity into account, except for [36], where transport to school is distinguished into 3 different modes: walking, biking, and motorized transport. Note that all studies define a certain study group, though some have tighter restrictions on their subjects. [35], [34], [32], and [33] examine adults, whereas [32] and [33] place additional requirements on these adults. The remaining two studies, [36] and [31], both study children of different age groups: 6 – 11 years and 11 – 18 years, respectively.

Query 3 The third query (Table 2) focuses on what the subjects are exposed to, dependent on whether the exposure is active (activity) or passive (environmental factor). In all active exposure cases, subjects are exposed to exactly the activities that are causing their exposure. For example, in [36], school children are exposed to walking, biking, or motorized transport. In [32] people are exposed to indoor fuel use (though indirectly via indoor air pollution), [34] subjects are exposed to eating at fast food outlets, and [35]’s subjects are exposed to physical activity. In the passive exposure studies, subjects are exposed to air pollution concentrations particulate matter 10, particulate matter 2.5, ozone, nitrogen dioxide, nitrogen oxides, carbon, and sulfur dioxide (PM10, PM2.5, O3, NO2, NOx, CO, and SO2) [33], and violent and non-violent crime [31].

Query 4 asks for the risk of exposure (Table 3). All studies identified some health-related risks as a consequence of the exposure. [31] focuses on mental health rather than physical health (risk of adverse behaviour). [36], [34], and [35] focus on obesity (though using different methods and considering different groups of people). [32] and [33]’s look at different risks of air pollution in their study, namely lung cancer and myocardial infarction and stroke.

Query 5 The first part of query 5 filters for environmental factors that influence exposure (Table 3). In the study for [33], concentrations for PM10, PM2.5, NO2, NOx, CO, and SO2 are identified as environmental factors. While many different types of chemicals are considered, the study lacks environmental factors from other environments such as distance to highways, that could also influence one’s exposure to air pollution. In [36], a wide range of envi-
environmental factors influence exposure: homes, schools, availability of major roads, distance to major roads, accident density, the proportion of cul-de-sacs, wind speed, temperature, global radiation, hourly precipitation, and proportion of green land use. Note that since [36] studies a form of active exposure, these factors directly influence physical activities, and only indirectly the exposure to those activities. [34] only considers two environmental factors, fast food outlets and neighbourhoods, while [35] focuses on green space and the built environment: large urban green space, quality of the urban green space, distance to green space, and distance to the city boundary. [31] takes into account only the social environment, including violent crime, non-violent crime, crime rates, and neighbourhood. [32]’s study focuses on coal deposits or mines and homes located in Chinese counties Xuanwei or Fuyuan are from the built environment, and socio-economic status is from the social environment. Note that in all passive exposure studies [31, 33], the involved persons are directly exposed to these environmental factors. The second part of query 5 asks about the data sets from which environmental factors were derived. In this query, there are many missing (None) answers because most studies provided only incomplete information on where data sets were obtained. As shown in Table 3, [36] provided links to data sources for wind speed, temperature, global radiation, hourly precipitation, and proportion of green land use. Our query also reveals that data on the availability of major roads, distance to major roads, and proportion of cul-de-sacs was derived from the same road data set. Similarly, accident density was derived from an accident data set. However, the data links to the road and accident data set are not available. Location data and other qualitative data on homes and schools were also not available (may be due to privacy reasons). [32] and [34] provided access to data about coal deposits or mines and fast food outlets, respectively. [31] and [35] did not provide any data sets for any of their environmental factors. Only [33] provided data links for all their environmental factors (monitoring stations).

Query 6 is about environmental stressors (Table 2). Answers for this query are lacking for all active exposures because they can never be caused by environmental stressors by definition (cf. Def 2 and 4). For example, large urban green space [35] is an environmental factor but is not a stressor. The only stressors, therefore, are air quality [33] and crime [31].

6. Discussion and future work

Ontologies are a way to organize knowledge in a field according to well-defined concepts. In combination with automatic annotation and information extraction methods, it can be used to handle large amounts of evidence on the influence of exposure in the health and behavioural sciences. In this article, we have focused on the design of an ontology that captures and makes comparable the conceptualizations of exposure and the underlying methods and data across different studies. The ontology categorizes parts of an epidemiological study in terms of the following related classes: person, activity, environment, exposure, dose, and risk. Using these classes as well as a universal causal relation, we defined different exposure concepts using OWL definitions. Based on OWL-RL/RDFS reasoning, we were able to categorize whether a given study in question focused on active and or passive exposure, which stressors are involved, who is exposed etc.

Our model illustrates the potential for an ontology to organize and extract information from exposure-related studies and classify it. It shows the variability of exposure conceptualizations including environmental causes and activities, but also basic commonalities, which allows us to compare articles against each other regarding their content. An article’s focus can be revealed by result frequencies (e.g., many environmental factors causing one activity, vs different activities in the same environment). Also, passive exposures tend to neglect activities and persons, whereas active exposures tend to model both. Many articles tend to lack information on data sets that were used to measure the exposure. The diversity of article samples encoded in our ontology shows that the ontology is general enough to cover various approaches to measuring exposure. This can be easily missed by keyword-based comparisons. For example, it is very easy to confuse Vermeulen’s [32] study about lung cancer and Lipsett’s [33] study about air pollution if we remain unaware that the former focuses on indoor fuel use (an activity), not on the quality of air. Yet both studies involve the keyword ”air pollution”. Thus our model is effective in adding semantic depth to meta-studies, which can now differentiate the underlying exposure model. Epidemiological researchers could use this to systematically compare approaches.
However, our work remains still preliminary in the following respects: One limiting factor of this study is that its empirical basis for testing is rather narrow, as only six articles were used to test the ontology. Such a small number was needed because each article needed to be read thoroughly and then encoded into the ontology and RDF manually, which is a very time-consuming, iterative process. So how could our study be scaled up to analyse larger amounts of articles? There is a potential to integrate natural language processing (NLP) and supervised classification into the framework to scale up the analysis of articles with our ontology. For example, a similar approach has been proposed by [4, 56]. However, while it can be assumed that e.g. BERT based pre-trained deep learning models [14] can classify text snippets as persons, activities, environmental factors etc with high quality, to date it remains unclear to what extent such methods are also able to extract the rather implicit causal relations between the categories investigated here and therefore remains a challenge [15, 54]. The latter would be needed to populate our ontology and to automatically infer different exposure categories. Furthermore, in the future, we should investigate to what extent the interpretation of ontology classes and relations are reproducible across different annotators based on measuring inter-annotator agreement [55]. In this way, we could find out to what extent our ontology classes allow for incompatible interpretations.

In addition, the ontology pattern that we proposed here can be improved in several ways. For one, the difference between passive and active exposure was modelled as a binary decision, however, it might be more adequate to allow for passiveness in degrees. For example, one could define an exposure as semi-active if all its causes are produced by some Activity, depending on the length of a chain of such causes that are routed in activities. This would allow us to recognize Vermeulen et al. [32]’s air quality study as an active case, even in case we conceptualized the indoor environment as an independent environmental factor. The causal chain would still reveal that such causes are all routed in the activity of burning coal, which can be controlled by the involved person. More generally, future work should investigate to what extent the used ontology of quantities [42] could be extended to capture the various ways how exposure measures are generated computationally. This would allow us to reason about the validity of method applications for certain measurement goals and, at the same time, to investigate the influence of systematic method variations on the quality of exposure-based models in the health sciences, and it could aid prediction models.

References

References


Appendix (selection of articles)

**Grinshteyn et al.** Grinshteyn et al.’s paper [31] studies how children’s mental health is impacted by witnessing, being a victim, or knowing a victim of violent or non-violent crime, specifically when the children show delinquent or aggressive behavior. Data was collected from seven cohorts in which children had been asked about their crime exposure. This data was then linked to uniform crime reporting data of the Federal Bureau of investigations (FBI). Based on this data, three sensitivity analyses were performed, with results showing the self-reported crime exposure was associated with increased scores [31].

**Vermeulen et al.** The paper by Vermeulen et al [32] investigates the relationship between lifelong exposure to the constituents of smoky coal and other fuel types, and lung cancer in females who do not smoke in two provinces in China. The researchers collected lung cancer cases among non-smoking women from six hospitals, and also used a control population. Both cases and controls were interviewed using a questionnaire that collected information on residential history, fuel use, and established or suspected risk factors for lung cancer [32]. Statistical analysis revealed that the strongest association with lung cancer was for a cluster of 25 polycyclic aromatic hydrocarbons (PAHS) and for NO2 [32]. This finding is in line with other studies but this was the first study known to examine the role of specific household air pollution constituents exposure of the entire life and lung cancer risk [32].

**Lipsett et al.** Lipsett et al [33] also look at air pollution, however from outdoors. The researchers’ goal was to examine associations of individualized long term exposures to particulate and gas usage air pollution with myocardial infarction and stroke in female teachers in California. This was done by linking geocoded addresses with inverse distance-weighting monthly pollutant surfaces for two measures of particulate matter and for several gaseous pollutants [33]. They examined associations between these pollutants and risks of myocardial infarction and stroke using Cox proportional hazard models [33]. Results showed long-term exposure to PM2.5, PM10 and NOx were associated with elevated risks for ischemic heart disease mortality [33].

**Van Rongen et al.** The study by Van Rongen et al [34] investigates Dutch neighborhood social norms with respect to fast food consumption as a potential mediating pathway between fast food outlet exposure and residents’ fast food consumption [34]. A sample of respondents living across the Netherlands completed a survey, where they reported on their fast food consumption and related perceived norms in their neighborhood [34]. The exposure to fast food...
was measured by the average count of fast food outlets within a 400m walking distance buffer around zip codes of respondents. Regression models were used to assess the association between residential fast food outlet exposure, fast food consumption, and social norm perceptions [34]. Results found that there was no overall direct association between residential fast food outlet exposure and residents’ fast food consumption [34]. However, the researchers found that fast food outlet exposure was positively associated with neighborhood social norms regarding fast food consumption, which was positively associated with the odds of consuming fast food [34].

**Hillsdon et al** Hillsdon et al’s [35] study examined the association between access to quality urban green space and levels of physical activity among adults living in Norwich, United Kingdom. This was done by performing three measures of access to open green space based on distance only, distance and size of green space, and distance, size, and quality of green space [35]. These measurements were done using GIS, and multiple regression models were used to determine relationships between the three factors and level of recreational physical activity. Results showed that there were no clear relationships. The authors concluded that access to urban green spaces does not appear to be associated with recreational physical activity for their sample group [35]. This article is interesting because it is the only article included in this study where no relationship was found between environmental factors with an action, exposure, or risk.

**Helbich et al** Lastly, Helbich et al’s [36] study is about measuring how the natural and built environment impacts Dutch children’s mode of transport to school, which may influence their exposure to physical activity, which in turn prevents obesity. This was done by giving children GPSs for several days, and by analysing the association between the environment on the school path and children’s active/passive transportation behaviour using mixed models. Results showed that distance to school, green space, and weather are not significant, but well connected streets and cycling lanes are [36].
<table>
<thead>
<tr>
<th>Paper</th>
<th>What exposures are modelled in this paper? (Query 1)</th>
<th>What types of exposures are these? (Query 1)</th>
<th>Which activities are involved in the exposure? (Query 2)</th>
<th>Who is exposed? (Query 2)</th>
<th>What subjects are exposed to? (Query 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helbich_2016</td>
<td>exposure to physical activity</td>
<td>expB:RiskPreventingExposure, expB:ActiveExposure</td>
<td>walking or biking or motorized transport</td>
<td>school children (GPS tracks)</td>
<td>walking or biking or motorized transport</td>
</tr>
<tr>
<td></td>
<td>PM 10 exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td>Living in California</td>
<td>female teacher</td>
<td>PM 10 concentration raster</td>
</tr>
<tr>
<td></td>
<td>PM 2.5 exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td>PM 2.5 concentration raster</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipsett_2011</td>
<td>O3 exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td>O3 concentration raster</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NO2 exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td>NO2 concentration raster</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NOx exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td>NOx concentration raster</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>CO exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td>CO concentration raster</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SO2 exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td>SO2 concentration raster</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vermeulen_2019</td>
<td>exposure to smokey coal</td>
<td>expB:RiskPromotingExposure, expB:ActiveExposure</td>
<td>indoor fuel use</td>
<td>never smoking women in the Chinese counties Xuanwei and Fuyuan</td>
<td>indoor fuel use</td>
</tr>
<tr>
<td></td>
<td>exposure to smokeless coal</td>
<td>expB:RiskPromotingExposure, expB:ActiveExposure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rongen_2020</td>
<td>poor diet</td>
<td>expB:RiskPromotingExposure, expB:ActiveExposure</td>
<td>eating at fast food outlets</td>
<td>adults in the Netherlands</td>
<td>eating at fast food outlets</td>
</tr>
<tr>
<td></td>
<td>witnessed violent crime exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td>living in crime neighborhoods</td>
<td>children aged 11 to 18 years old</td>
<td>violent crime</td>
</tr>
<tr>
<td>Grinshtein_2018</td>
<td>hearsay violent crime exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td></td>
<td></td>
<td>non-violent crime</td>
</tr>
<tr>
<td></td>
<td>victim of violent crime exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>witnessed non-violent crime exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hearsay non-violent crime exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>victim of non-violent crime exposure</td>
<td>expB:PassiveExposure, expB:RiskPromotingExposure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hillsdon_2006</td>
<td>exposure to physical activity</td>
<td>expB:RiskPreventingExposure, expB:ActiveExposure</td>
<td>physical activity</td>
<td>adults in Norwich, England</td>
<td>physical activity</td>
</tr>
</tbody>
</table>

Table 2
Answers to queries 1-3 retrieved from the knowledge base via inference.
### Table 3

<table>
<thead>
<tr>
<th>Paper</th>
<th>What is the risk of exposure? (Query 4)</th>
<th>Which environmental factors influence the exposure? (Query 5)</th>
<th>From which datasets were they derived? (Query 5)</th>
<th>What are the environmental stressors? (Query 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helbich_2016</td>
<td>obesity</td>
<td>homes</td>
<td>None</td>
<td>roads, None</td>
</tr>
<tr>
<td></td>
<td></td>
<td>schools</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>availability of major roads</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>distance 2 major roads</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>accident density</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Lipsett_2011</td>
<td>Myocardial Infarction</td>
<td>PM 10 concentration raster</td>
<td>PM 10 monitoring stations, <a href="https://www.arb.ca.gov/adam">https://www.arb.ca.gov/adam</a></td>
<td>PM 10 concentration raster</td>
</tr>
<tr>
<td></td>
<td>Stroke</td>
<td>PM 25 concentration raster</td>
<td>PM 2.5 monitoring stations, <a href="https://www.arb.ca.gov/adam">https://www.arb.ca.gov/adam</a></td>
<td>PM 25 concentration raster</td>
</tr>
<tr>
<td></td>
<td></td>
<td>O3 concentration raster</td>
<td>O3 monitoring stations, <a href="https://www.arb.ca.gov/adam">https://www.arb.ca.gov/adam</a></td>
<td>O3 concentration raster</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NO2 concentration raster</td>
<td>NO2 monitoring stations, <a href="https://www.arb.ca.gov/adam">https://www.arb.ca.gov/adam</a></td>
<td>NO2 concentration raster</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NOx concentration raster</td>
<td>NOx monitoring stations, <a href="https://www.arb.ca.gov/adam">https://www.arb.ca.gov/adam</a></td>
<td>NOx concentration raster</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CO concentration raster</td>
<td>CO monitoring stations, <a href="https://www.arb.ca.gov/adam">https://www.arb.ca.gov/adam</a></td>
<td>CO concentration raster</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SO2 concentration raster</td>
<td>SO monitoring stations, <a href="https://www.arb.ca.gov/adam">https://www.arb.ca.gov/adam</a></td>
<td>SO2 concentration raster</td>
</tr>
<tr>
<td>Vermeulen_2019</td>
<td>lung cancer</td>
<td>coal deposits or mines</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>homes located in Chinese Counties Xuanwei or Fuyuan</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>socio-economic status</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>household characteristics</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Rongen_2020</td>
<td>obesity</td>
<td>fast food outlets neighbour hood</td>
<td><a href="https://locatus.com/applicatie/retail-facts/">https://locatus.com/applicatie/retail-facts/</a></td>
<td></td>
</tr>
<tr>
<td>Grinshteyn_2018</td>
<td>adverse behavioural health characteristics</td>
<td>violent crime</td>
<td>None</td>
<td>Violent crime</td>
</tr>
<tr>
<td></td>
<td></td>
<td>non-violent crime neighbour hood</td>
<td>None</td>
<td>Non-violent crime</td>
</tr>
<tr>
<td></td>
<td></td>
<td>crime rates</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Hillsdon_2006</td>
<td>obesity health issues</td>
<td>large urban green space quality urban green space distance to green space distance to city boundary</td>
<td>None</td>
<td></td>
</tr>
</tbody>
</table>

Answers to queries 4-6 retrieved from the knowledge base via inference.