Context-Based Ontology for Urban Data Integration

Michal Med[1] and Petr Křemen[2]

[1] Czech Technical University in Prague, Czech Republic
TODO@fel.cvut.cz
[2] Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
petr.kremen@fel.cvut.cz

Abstract. Urban planning data are of big importance to both general public and domain experts like civil engineers, architects or urban planning specialists. Typically, the data are scattered within many datasets containing both geographical knowledge and taxonomical knowledge. The distributed nature of such data, as well as their complexity practically prevents formulating relevant geographic queries over multiple datasets.
In this paper, we present a case study on ontology modeling of urban planning data of the City of Prague. We discuss the ontological nature of the domain, as well as formalization of the ontology in the OWL language. At the end of the paper we present various distributed queries over multiple datasets.

1 Introduction

Urban planning is a data-intensive domain encountering great interest of experts as well as of general public. Many data sources in urban planning are public by their nature, yet, different level of quality of the data sources, as well as their semantic heterogenity makes them difficult to interpret and exploit. Although ontologies for urban planning have been developed for over a decade (e.g. [1], [2]), not much attention has been paid to the context-based knowledge representation in the domain.

We came accross the need of context-sensitive ontology-based integration model during our research project aimed at integrating and exploiting urban data sources of the Prague Institute of Planning and Development (IPR). The institute has tens of datasets, content of which is typically visualized by means of a GIS. However, the integration of their content is missing. It turns out that many generic terms, like *building*, *construction*, or *house*, are understood differently by architects, civil engineers, urban planning experts as well as general public. These semantic discrepancies have significant impact on subsequent data usage including statistics. Also, even a technique that would provide a limited set of datasets that would be necessary to explore manually in order to evaluate given query is missing.

Here we introduce a *context-sensitive ontology-based model* of urban planning dataset integration. The model is based on the UFO ontology [3] and its extensions by power types [4]. The model is then formalized in an OWL 2 [5] ontology using SPARQL [6], resp. $SPARQL - DL^{NOT}$ [7] queries for distributed query formulation. The queries are discussed in terms of usability in the domain.

Section 2 presents the most important notions used in the paper.

2 Background

2.1 Unified Foundational Ontology

As a top-level ontology, we use the Unified Foundational Ontology (UFO). Comparing to other approaches (e.g. [8], [9], [10]), it is an actively developed ontology that incorporates structural modeling, event modeling, mereology, as well as power types and context modeling. Furthemore, it provides an ontology design UML-based language OntoUML.

UFO consists of:

UFO-A is a foundational ontology analyzing structural modeling constructs [11], describing object types, taxonomies, associations and mereological relations,

UFO-B is an ontology of events [12], describing events and situations

UFO-C is an ontology of social and intentional aspects [13], describing intentions and social aspects of agents, and

UFO-S is an ontology of services [14], describing service agreements and commitments.

UFO-A defines fundamental ontological categories, making distinction between *individuals* (e.g. one particular person) and *universals* (e.g. a type *person*), representing types of individuals. Next fundamental distinction on individuals is between *endurants* (e.g. a person) and *perdurants* (e.g. an event). Endurants can be observed as complete concepts in a given time snapshot, while perdurants only partially exist in a given time snapshot. Endurants can be e.g. *objects* (an apple), or its *tropes* (e.g. color of an apple).

UFO-B extends concepts from UFO-A with the notion of an *event*, a perduring entity, spanning some time interval, or occuring in time instant and having participants and parts (sub-events). Events themselves can be temporally related, i.e. one hapenning before another, during another, etc. as specified by Allen's temporal algebra [15]. A *situation*, on the other hand, can be understood as a bag of snapshots of object states and their relationship at one particular time instant. Consequently, an event has its presituation (composed of object snapshots valid just before the event started), and a post situation (composed of object snapshots valid just after the event ended). Such framework is suitable for representing temporal and causal relationships between events, as well as for simulation scenarios [12].

UFO-C extends UFO-B with notions of *agents*, i.e. proactive objects with intention (e.g. a person, or a company), their intentions, *commitments* and *actions* they perform.

UFO-S extends UFO-B with the definition of a *service*, service agreements, requirements and obligations by service producers and consumers.

Main parts of the UFO ontology that are fundamental for the purpose of this paper are depicted in Figure 1.

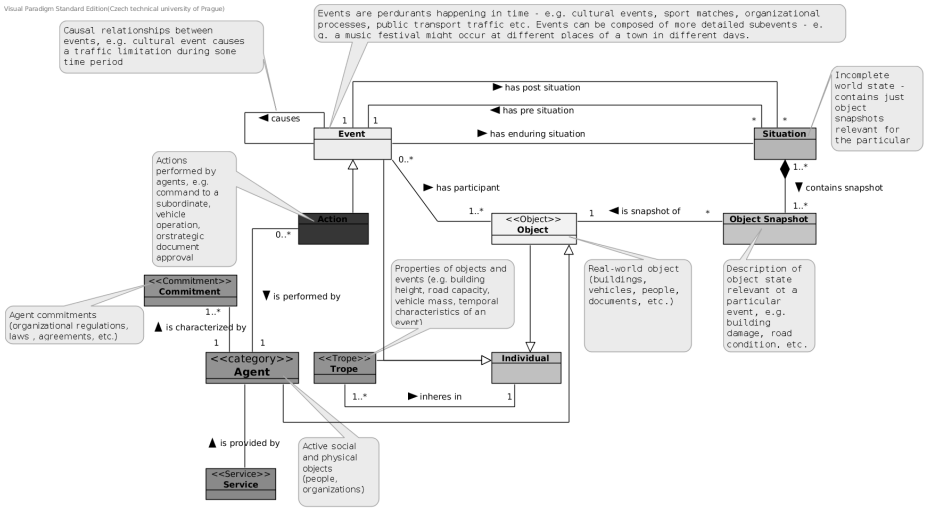

Fig. 1. Fundamental concepts and relations of UFO.

OntoUML OntoUML is a language provided on top of UFO for the purpose of conceptual modeling. The language is introduced in [11]. We will review the most important notions necessary for the purpose of this paper.

Sortal vs. Mixin. a *Sortal* is any universal carrying a principle of identity, e.g. *Person* universal (identifiable by DNA, or birth number, etc.), contrary to *Mixin* universal, e.g. *Object* universal having different principles of identity given by its sub types - Person, Building, Vehicle, etc.

Rigid vs. Anti-rigid. *Rigid* is any universal, for which all its instances (individuals) belong to the universal for the whole time of their existence, e.g. *Person* sortal universal. *Anti-rigid* is any universal for which all its instances will not be in the future, or were not in the past instances of the universal, e.g. *Moving Vehicle* sortal universal.

Kinds and subkinds. a *Kind* is any rigid sortal, that provides principle of identity, e.g *Person* kind, contrary to a *subkind*, any rigid sortal principle of identity of which is inherited from its ancestor *Kind*, e.g. *Man* subkind inherits its identity from the *Person* kind.

Phase sortals and role sortals. a *Phase sortal* is an anti-rigid sortal for which intrinsic properties of an individual are those that make the individual an instance of that anti-rigid sortal, e.g. being a *Teenager* depends on the age of the particular person. In contrast, a *Role sortal* is an anti-rigid sortal for which relation with another individual makes the particular individual an instance of that anti-rigid sortal, e.g. being a *School building* depends on the use of the particular building.

Relators and relations. A *Relator* reifies the notion of a relation among several individuals, e.g. *Construction* (as a relation between a company, a building, etc. existing during the *Construction* event in given time frame).

Powertypes. A powertype represents the notion of a *type of types* introduced and analysed in [16]. Powertypes is a fundamental notion for product type modeling (like *building type*), or social role modeling (like *Urban planning expert type*).

3 Urban Planning Ontology

The whole process of ontology modelling has been conducted for over ten months and can be separated in several phases. In the first phase, ontology designers became familiar with data published IPR and chose the adequate ontology modeling technique. In the second phase, ontology model prototypes based on the basic knowledge of data were created and consulted with data professionals employed by IPR. They have provided information and annotations of five thematically similar datasets. Third phase of design started with ontological modelling of datasets. Because of terminology ambiguities, most of the notions are contextualized Data are put in contexts and therefore, single ontologies for every context were created.

3.1 Data analysis

Prague Institute of Planning and Development manages tens of datasets, publishing them as open data on their open data geoportal. Datasets contain data related to the city management, urban planning data, data about traffic, POIs and more. For the publication, standardised ATOM service is used. All data sets shall be provided with metadata according to the ISO 19115[17]. Data are discovered on the geoportal as well, by the keywords. Set of keywords, including names of spatial objects, their attributes and words used for searching data by users themselves, was provided by IPR. The list contains more than 200 keywords with different meanings. Some of the concepts are used in various meanings, some concepts have the same meaning as another, e.g. `park and ride`, `Park & ride` and `P + R`. Some concepts may have different meaning in different contexts, e.g. *construction* and *building* may have similar or same meaning in the context of the Land use dataset, but in the context of the Buildings dataset, *building* is a specialization[] of a *construction*, i.e. every *building* is a *construction* and all properties of a *construction* are properties of a *building* as well. Moreover,

concept of *construction* can be used in the meaning of a process or an object. Context differs among datasets, specific groups of users, object or attribute type and more.

First task was to sort concepts, find one common theme and try to find relations between concepts related to the chosen theme. According to the first look on the set of concept, traffic seemed to be suitable theme to start with. First conceptual model was created on the basis of railway traffic. For the first model, mainly specializations were used and only few object properties. Fourteen concepts related to the railway were derived from the set and the contexts and relations were searched for among them. For example, concepts *railway*, *road* and *rail crossing* are definitely somehow related. Relations between objects are ambiguous and for the person without any knowledge and expectations, it is quite difficult to discover such ambiguities. In the example above, *crossing* may be part of *railway*, but also, it may be the part of *road*. *Road* and *railway* may touch, but their connection may be done only in *crossing*. In some cases, *railway* could have been removed, but *crossing* is still on its place. Inexperienced mind also tends to think about same objects in different meanings in different situations. It was proved that the most important thing is to capture meanings for every concept properly and document it (by means of an OWL annotation), and then count every meaning as one object. Therefore it is very important to have at least common knowledge of the topic that is described by the model.

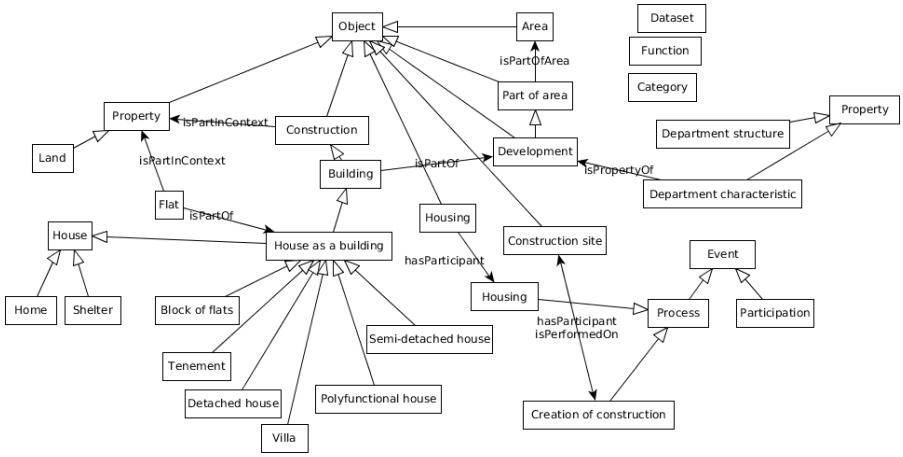

Fig. 2. The first semantic model of the ontology created over urban planning data.

The analysis of railway traffic was performed in order to become familiar with the domain and setup basic notions for urban planning. In the next steps we extended the model of traffic in Prague with more interlinked data in order to test the potential to answer distributed queries. The issue of urban planning

is still quite wide for the basic model. For this purpose the domain of *housing* was selected by domain experts from IPR.

3.2 Ontology modelling

The first model of urban planning ontology was designed from bottom to the top – first, basic concepts for the housing issue were derived from the set into the new ontology. Second step was to find and mark relations between objects in the model and important relations heading out of the existing model. Relations and outer objects were added to the model. Besides specializations, object properties starting to be more important in the conceptual model of housing (see Figure 2). Still, every object can be seen in various roles depending on the context. At this stage, the crucial problem was to correctly define meanings (concepts) for different terms in different contexts.

Very important input to the model came from IPR, because the meaning of some concepts were misunderstood. Good example can be the word *real estate*, having different meaning in the context of various groups of users – for urban planners it is used as *unmovable object, including flats*, for brokers it is *a building or ground, but also a flat or even movable objects such as residential caravans*, and from the perspective of an employee of the cadastral office, *building may be a part of ground and some juridical relations are considered real estates*. These records had to be set straight.

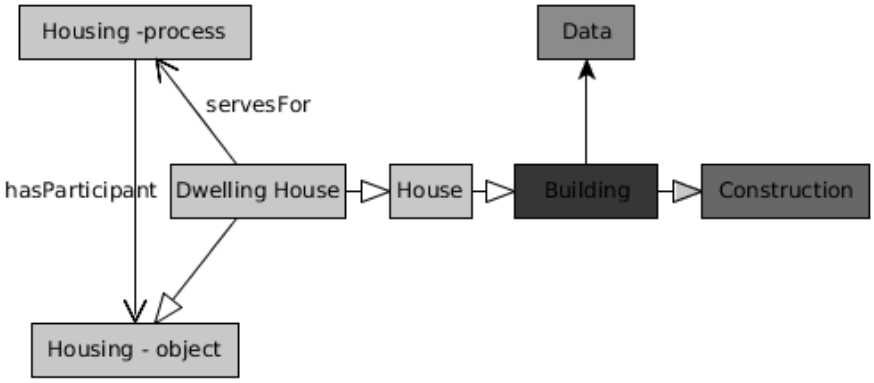

Fig. 3. Model of buildings according to the technical map.

At this stage of work, design split into integrated thematic units. Ontology was planned to be used in the context of datasets managed by IPR, therefore it was necessary to understand structure and content of those datasets. As a content of the sample set of datasets, five datasets were chosen:

Buildings is a dataset containing polygons of buildings according to the technical map,
Floors is a dataset containing data about buildings with information of their heights and number of floors,
Current land use is a dataset containing data describing areas and their parts, their function and form of development in the area,
Functional land use is a dataset, that is part of spatial plan, containing data describing designed functional land use,
Parcels is a dataset containing data about cadastral parcels.

For each dataset, visual model of relations and involved classes was created. The result is in Figures 3, 4, 5, 6 and 7. In the models, the darkness of the class is proportional to the importance of the concept w.r.t. the particular dataset.

Buildings In the dataset Buildings there are almost the same information as in the dataset Floors. Differences are in attributes and in the context. Buildings are not in wide context, therefore it is the simplest dataset among the five chosen, at least according to the number of related classes. The main class in the dataset is building, carrying information about it in the form of individuals:

identifier is a unique code for every building,
house number is identifier of building in a given area,
house number type is a house type identifier.

Building object also carry information about lifespan and publisher.

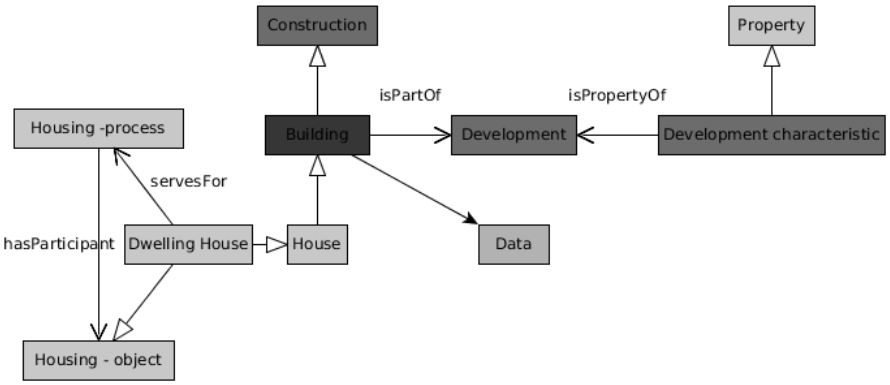

Fig. 4. Model of buildings and objects relative to them and their properties in the context of floors dataset.

Floors Floors shall contain information about heights and shape of buildings. Shape of buildings in the area affects the characteristics of development. Therefore, more classes are possibly related to this theme and model is wider, although

it contains information about the same kind of object. Attributes carried by building class are:

type of object is the type of building, value itself comes from the codelist,
number of floors is the sum of floors both above and under ground,
type of roof defines the shape of roof,
number of roof floors defines how many floors are narrowed by the roof,
number of floors in the slope defines how many floors may be partially underground.

Building object also carry information about lifespan and publisher.

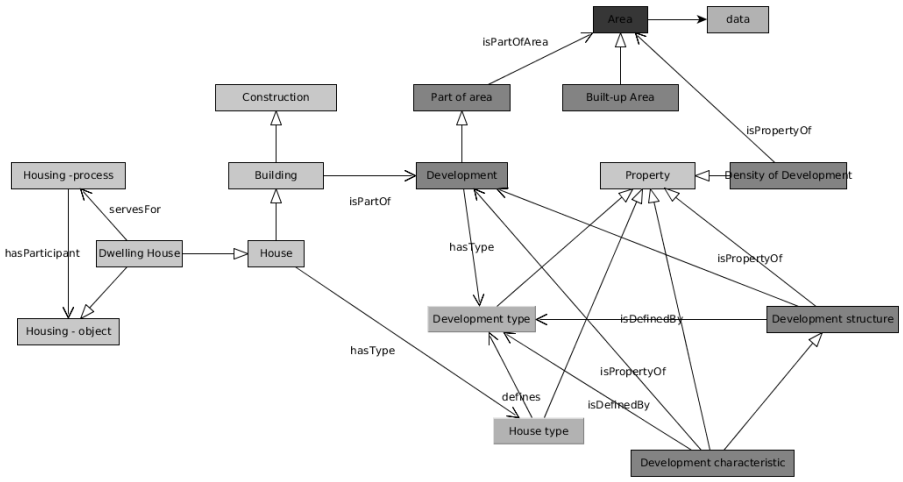

Fig. 5. Model of current land use relative obects and their properties.

Current land use Information–bearing object is an area. Objects in the context of this datasets carries information about the real land use. It may differ from the planed or purposed land use. Context covers really lot of objects, because current land use is influenced by lot of aspects.

Information carried by area object itself is not so rich:

land use code is the information about main current land use, value itself comes from the codelist,
behind Prague defines, if the area lies within the capital borders, or not,
other use defines secondary land use, value itself comes from the codelist.

Area object also carries information about publisher.

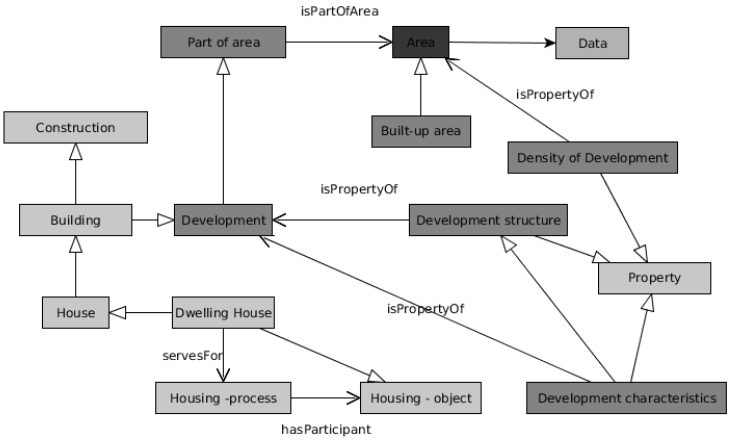

Fig. 6. Model of functional land use relative objects and their properties.

Functional land use Information–bearing object in the dataset is an area. In the urban planning, areas have attached proposed usage. Some of areas are not used the way they should according to this information. Part of development characteristic may be the shape of buildings on it as well. Context of relations in this dataset is quite wide – a lot of properties are influencing the land use, but not as many as in the case of current use. In the planning, not all possibilities are considered. In the design of functional use, two values are filled for each attribute – one for design horizon and second one as territorial reserve:

functional use of area code is the value from codelist describing planned use of the area,
code of the land use rate is a value from codelist,
number differentiation is a number from 1 to 7,
composite code is a text.

Parcels Dataset parcels contains information about cadastral parcels and their affiliation to the cadastral zonings. For the purpose of the model design of this dataset, new class had to be created. The class is called cadastral zoning and in the context of urban planning, its meaning is relevant only for this dataset (as far as is the context of five designed datasets). The main information–bearing class is parcel, but cadastral zoning carries on attribute too. Attributes of parcel are:

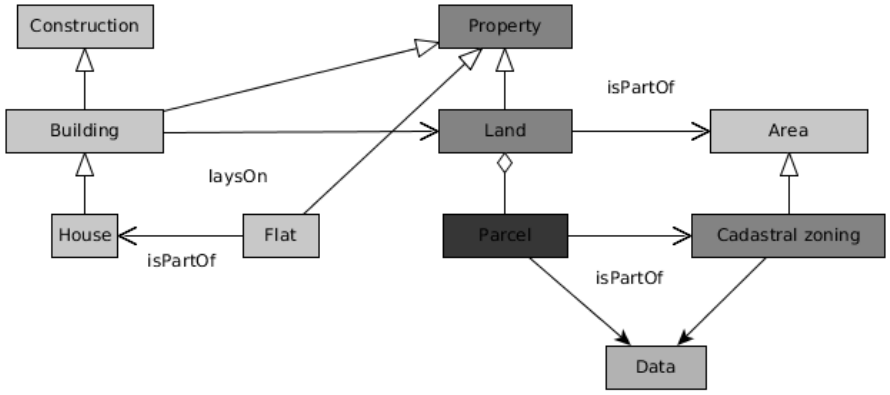

Fig. 7. Model of cadastral parcels and objects relative to them and their properties in the context of parcels dataset.

identifier is a unique code within whole dataset,
parcel number is the unique identifier within cadastral zoning,
date of creation/last change/deletion is the par of the objects life cycle,
area is the acreage of the parcel in square meters.

Cadastral zoning has only one attribute:

code of the cadastral zoning is unique identifier of cadastral zoning within whole dataset.

The word *Property* in the context of dataset Parcels has different meaning than *Property* in the context of datasets Current land use and Functional land use. In the previous cases, property is an abstract class pointing to an attribute or feature (e.g. density is property of population). In the context of parcels, property is an object owned or possessed by some entity, private or legal (e.g. building is property). Interesting fact is, that this problem depends on the language used in model. In Czech, two meanings of property are differently named.

All models, including classes (objects) and properties (relations) were transposed into the ontology. In datasets is a big amount of concepts used in two or more of datasets. Some of them are used in different contexts in every occasion. At this stage, dividing classes into more is not the simplest way of adding context to the ontology. Since the contexts of concepts begun to be more important to the completeness and entirety of the model, more effective way of distinguishing contexts had to be found. It was decided to divide ontologies into more files, according to its context.

3.3 Semantic contexts

For the purposes of this work, context is represented by datasets. For every dataset, a single ontology was created, represented by single file. For every

dataset described in 3.2, new ontology was created. In those ontologies are held only the properties, classes and individuals related to the context of dataset. Classes and properties having the same ontological meaning in every use are stored in the ontology *common*. The rest of the concepts, that were not used in any of datasets were moved to the ontology *appendix*. In the ontology common, only common objects and relations are visible. In ontologies representing dataset contexts, all content from *common* ontology are visible as well. For the better information about context of dataset, every class, individual and property is annotated as *isInContextOfDataset*. For the visualisation and querying over the whole urban planning ontology, ontology *all* was created. Summary of ontologies and relations among them is shown in the Figure 8. Some of the properties may be used in the same context even in more than one ontology. For these cases, more ontologies can be created and then imported into all applicable ontologies (bright blue).

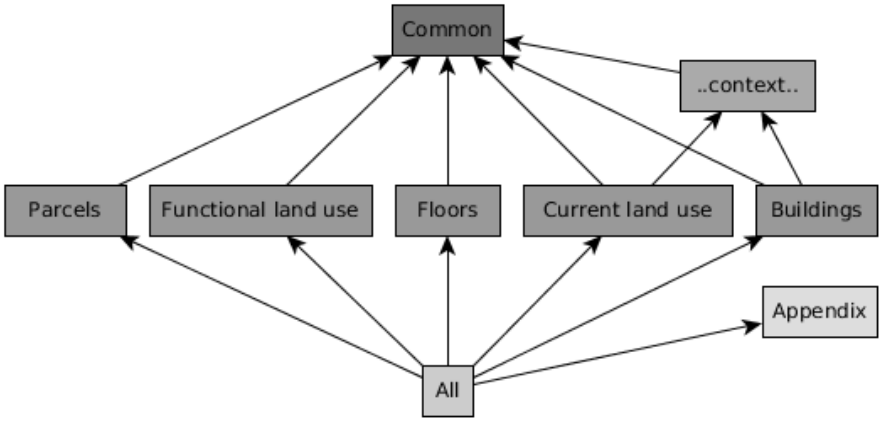

Fig. 8. Summary of dataset ontologies and relations among them.

Context itself is defined as a class in the *common* ontology as a specialization of information object. Possibility of adding new types of context in the future remains open, therefore the structure of information objects remains more branched, than needed on the first sight.

4 Practical Queries and Evaluation

The purpose of ontology model is to get rich linked information from the datasets. One of the proposed targets is creation of a querying mechanism over the IPR datasets on their geoportal. Their idea is querying mechanism, that returns not only exact matches, but rates related responses and returns matches, that may

have something in common. Problem of browsing data by users is, by the words of IPR, that users do not know, what exactly do they expect. It is quite import to know, what kind of queries IPR expects to be asked. They have provided us with following examples:

1. What areas shall be used for housing according to the spatial plan, but their current use is different and are owned by municipality?
2. In which areas lie apartment houses taller than four floors?
3. How many percent of the floor area in the Old Town is used for housing?
4. How many family houses are in Uhříněves?
5. Which office buildings have no parks in their neighbourhood?
6. What is the floor area of all office buildings in Prague?
7. Which gas stations lie in areas, where no gas stations shall lie?
8. In which municipal areas are tenement houses?
9. Which buildings in proximity of given point are owned by public subjects and are able to be used for the building of new library?

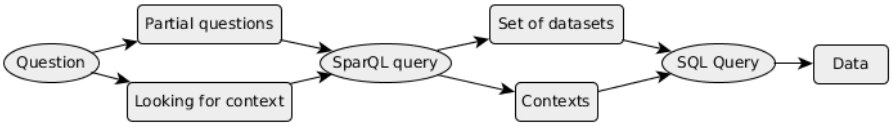

Fig. 9. Process of asking question over the ontology and datasets.

All of these questions are quite complex and lot of them may be answered in different ways. All of these questions are querying data itself. Data are not part of ontology model. Moreover, some concepts are not clear in its meaning, e.g. Uhříněves can be a name of part of municipality or cadastral zoning. Another question is on the meaning of words like "neighbourhood", "shall", "used for" and similar. Good example of many meanings in one question is number eight: In which municipal areas are tenement houses? Municipal area may be administrative unit of municipality, but it also may be an area inside municipality or area of parcels, on which is urban area. Second part of the question is problematic as well. Answer may refer to houses with the type "tenement", or houses in areas with attribute "used for tenement". Each question shall be divided into elements and for this elements context (or more possible contexts) shall be found. For the search, ontology querying shall be used.

For the querying the context and structure of contents used in question is used SparQL language. Queries are applied upon the *all* ontology, that contain all information about datasets, including context. Questions above could be split into set of SparQL queries defining context and datasets, where to look for information. The result of SparQL shall be the set of datasets and classes in specific context, relevant for finding the answer on asked question.

Basic set of SparQL queries shall provide answer on following questions:

1. in which datasets can be found given concept?
2. in which context is concept used over the set of datasets?
3. in which dataset can be given found attribute value?
4. what properties are there between two concepts in a given set of datasets?
5. which datasets contain information needed to answer question on specific data?

As seen before, first four questions are basically subsets of the fifth. Question number five is the most important question and in the practical use it shall be used for searching context for further SQL query focused on direct access to data.

The process of querying proceeds as follows (see Figure 9):

1. User asks a question on the IPR geoportal (that will be probably provided by the third party of the project, private company Gisat),
2. question is split into parts,
3. context of single parts is queried by SparQL queries,
4. output for the queries is usage and meaning of concepts in proposed context, including datasets and their structures, where specific data can be found,
5. based on the context, SQL queries are created and commited (this will be provided by Gisat).

5 Conclusions

In the first year of project, most of the work was done on the ontological models. In the beginning, it was important to get used to the structure and meaning of data and nomenclature used in the field of urban planning. After the first tries of creating ontologies, few specifics were found. It is very important to not underestimate annotations of classes and properties. As the time passes and the number of concepts raises, more and more properties depends on context.

Acknowledgements This work was supported by the grant No. TA04021499 "Open Data and semantic approaches to uncover social aspects of urban quality" of the Technology Agency of the Czech Republic.

References

1. Frederico T Fonseca, Max J Egenhofer, CA Davis, and Karla AV Borges. Ontologies and knowledge sharing in urban gis. *Computers, Environment and Urban Systems*, 24(3):251–272, 2000.
2. Achilleas Psyllidis. Ontology-based data integration from heterogeneous urban systems: A knowledge representation framework for smart cities. In *CUPUM 2014: Proceedings of the 14th International Conference on Computers in Urban Planning and Urban Management, Cambrigde, USA, 7-10 July 2015*. MIT, 2015.

3. Giancarlo Guizzardi. *Ontological Foundations for Structural Conceptual Models*. Number 15 in Telematica Institute Fundamental Research Series. Telematica Instituut, Enschede, The Netherlands, 2005.
4. Giancarlo Guizzardi, Almeida Joo Paulo A., Guarino Nicola, and Carvalho Victorio A. Towards an ontological analysis of powertypes. In *JOWO 2015: The Joint Ontology Workshops co-located with IJCAI 2015*, 2015.
5. W3C OWL Working Group. *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation, 27 October 2009. Available at http://www.w3.org/TR/owl2-overview/.
6. SPARQL 1.1 Query Language. Technical report, W3C, 2013.
7. Petr Kremen and Bogdan Kostov. Expressive owl queries: Design, evaluation, visualization. *Int. J. Semant. Web Inf. Syst.*, 8(4):57–79, October 2012.
8. Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. Sweetening ontologies with dolce. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, EKAW '02, pages 166–181, London, UK, UK, 2002. Springer-Verlag.
9. Pierre Grenon and Barry Smith. Snap and span: Towards dynamic spatial ontology. *Spatial Cognition & Computation*, 4(1):69 – 104, 2004.
10. Ian Niles and Adam Pease. Toward a Standard Upper Ontology. In Chris Welty and Barry Smith, editors, *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9, 2001.
11. G. Guizzardi. *Ontological foundations for structural conceptual models*. PhD thesis, University of Twente, Enschede, The Netherlands, Enschede, October 2005.
12. Giancarlo Guizzardi, Gerd Wagner, Ricardo de Almeida Falbo, Renata S. S. Guizzardi, and Joo Paulo A. Almeida. Towards ontological foundations for the conceptual modeling of events. In Wilfred Ng, Veda C. Storey, and Juan Trujillo, editors, *ER*, volume 8217 of *Lecture Notes in Computer Science*, pages 327–341. Springer, 2013.
13. Peter Green and Michael Rosemann. *Business Systems Analysis with Ontologies*. Idea Group Publishing, June 2005.
14. Julio Cesar Nardi, Ricardo de Almeida Falbo, Joo Paulo A. Almeida, Giancarlo Guizzardi, Lus Ferreira Pires, Marten van Sinderen, and Nicola Guarino. Towards a commitment-based reference ontology for services. In Dragan Gasevic, Marek Hatala, Hamid R. Motahari Nezhad, and Manfred Reichert, editors, *EDOC*, pages 175–184. IEEE Computer Society, 2013.
15. James F. Allen. Maintaining knowledge about temporal intervals. *Commun. ACM*, 26(11):832–843, November 1983.
16. Giancarlo Guizzardi, Joao Paulo, Paulo Joao Almeida, Nicola Guarino, and Victorio Albani Carvalho. Towards an ontological analysis of powertypes. In *Proceedings of the International Workshop on Formal Ontologies for Artificial Intelligence (FOFAI 2015)*, 2015.
17. ISO. Geographic information – metadata – part 1: Fundamentals, 2014.