DC Proposal: Towards a Framework for Efficient Query Answering and Integration of Geospatial Data*

Patrik Schneider
Advisors: Thomas Eiter and Thomas Krennwallner

Institut für Informationssysteme, Technische Universität Wien
Favoritenstraße 9-11, A-1040 Vienna, Austria
patrik@kr.tuwien.ac.at

Abstract. Semantic Web technologies are becoming more interleaved with geospatial databases, which should lead to an easier integration and querying of spatial data. This is fostered by a growing amount of publicly available geospatial data like OpenStreetMap. However, the integration can lead to geographic inconsistencies when combining multiple knowledge bases. Having the integration in place, users might not just issue a points-of-interest search, but rather might be interested in regions with specific attributes assigned to them. Though, having large amounts of spatial data available, standard databases and reasoners do not provide the means for (quantitative) spatial queries, or struggle to answer them efficiently. We seek to combine spatial reasoning, (nonmonotonic) logic programming, and ontologies for integrating geospatial databases with Semantic Web technologies. The focus of our investigation will be on a modular design, on efficient processing of large amounts of spatial data, and on enabling default reasoning. We propose a two-tier design related to HEX-programs, which should lead to a plausible trade-off between modularity and efficiency. Furthermore, we consider suitable geo-ontologies to semantically annotate and link different sources. Finally, the findings should lead to a proof-of-concept implementation, which will be tested for efficiency and modularity in artificial and real-world use cases.

1 Background and Problem Statement

Fostered by a popular demand for location-aware search applications, linking and querying spatial data has become an active research field. At the same time, governments open up their official datasets for public use, and collaborative projects like OpenStreetMap (OSM) are becoming large sources of spatial data (http://www.openstreetmap.org/). In this context, geospatial databases are the backbone for storing and querying these data. Hence they have been extensively studied by the Geographic Information Systems (GIS) community (cf. [9, 18]).

Geospatial databases often have the drawback that querying them is complicated, inference mechanisms are virtually non-existent, and extending them is difficult. In response, Semantic Web technologies are becoming more interleaved with geospatial databases [5, 24], which should lead to an easier integration and querying of spatial

* Supported by the Austrian Research Promotion Agency (FFG) project P828897, the Marie Curie action IRSES under Grant No. 24761 (Net2), and the EC project OntoRule (IST-2009-231875).

data. The integration should happen on several levels. First, the different data sources have to be linked to concepts and roles of different (often predefined) ontologies. After having several ontologies and their assertions in place, they have to be merged or linked in a certain manner. However, on the spatial level, linking can cause geographic inconsistencies, e.g., if the same place is located on different coordinates because of imprecise data. On the logical level, inconsistencies arise by introducing contradictions in the joint knowledge base (KB). Searching and retrieving location-based information is possible with a working integration. But users might be interested in *areas* with specific attributes assigned to them instead of searching plain or semantically annotated points-of-interest (POIs). For example, the walk-ability of a certain neighborhood in a city could be of interest (cf. http://www.walkscore.com/). However, having spatial data of larger cities or even countries, standard tableaux-based reasoners do not provide the means for (quantitative) spatial queries or struggle to answer them efficiently.

We seek to combine spatial reasoning, (nonmonotonic) logic programming rules (such as Answer Set Programming [6], Prolog [2], or Semantic Web Rule Language (SWRL) [17]), and ontologies for interleaving the different approaches. Besides having rules to formalize spatial integrity constraints [1], rules opens a way to qualitative spatial reasoning with the Region Connection Calculus (RCC) being applied on top of ontologies and spatial data [15]. Several authors have considered these combinations. We distinguish between heterogeneous and homogeneous combinations, where heterogeneous combinations can be separated into loose couplings and tight integrations. Grütter *et al.* focus on enhancing ontologies with spatial reasoning based on RCC and SWRL-rules to capture dependencies between different administrative regions [15]. In [1], a combined framework based on Description Logic Programs (DLP) was introduced. This approach considers the coupling with a spatial databases and the focus on the formulation of spatial constrains. With PelletSpatial and DLMAPS, there has been proof-of-concept implementations of qualitative spatial reasoning in a Description Logics (DL) reasoner, featuring consistency checking and spatial query answering [27, 28].

Modularity as a design goal enables the integration of external computation sources, which could be a DL reasoner, spatial data sources, or computational geometry engines. With efficiency as a goal, we have to identify efficient external computation sources, optimize information flow between them, and prune intermediate results. We need to put a particular focus on the trade-off between expressivity and performance first, and between pre-computation and on-demand calculation of queries thereafter. We will try to capture nonmonotonic notions such as reasoning by default to express exceptions. For example, it is important to express statements such as "by default all restaurants in a city are non-smoking," but we like to state some exceptions with a smoking permission. Concluding from the points above, we will focus on the following research questions:

- Given the focus of our investigation on modular design, efficiency, and nonmonotonic reasoning, which will be the most suitable architecture for a framework of combining spatial data, rules, and ontologies?
- What methods are feasible to infer certain regions from a selected point set? And, how to deduce attributes from the created regions? For example, we would like to investigate which point sets (e.g., cafés, restaurants, or pubs) make up a suitable neighborhood for dining.

– Considering several spatial data sources like OSM or Open Government Data (OGD), how can we combine this sources using logic programs as an integration mechanism?
– How does the framework behave for query answering over large data sets (e.g., the OSM data set of Austria)? And how does it perform in an artificial test environment and further under real-life conditions in an e-government and public transit system.

2 Related Work

A well known top level ontology is GeoOWL, which keeps a strict distinction between the geographic object, called *Feature*, and its footprint, called *Geometry* (http://www.w3.org/2005/Incubator/geo/XGR-geo/). Furthermore, GeoNames is a feature-centric geographical database containing about 7.5 million unique features (http://www.geonames.org/). In particular the categorization in nine top-feature (such as area feature, road feature, building feature, etc.) and corresponding sub-features (e.g., street, railroad, trail) are of interest. The creators of GeoNames also created an OWL ontology, which is very instance-heavy, with just a few concepts defined. Coming from the field of pervasive computing, the Standard Ontology for Ubiquitous and Pervasive Applications (SOUPA) is a general ontology, which covers the domain of space, time, actions, and agents [8]. The top concept *SpatialThing* is divided into the sub-concepts *GeographicalSpace* and *RCCSpatialRegion*. Different from other frameworks, the RCC calculus is considered as a concept instead of a set of transformation rules.

RCC8 is a fragment of RCC, where eight binary predicates are defined for representing the relationships between two regions [3, 23]. Drawing from the close connection between DL and Modal Logics, and between *RCC8* and Modal Logics, Katz and Cuenca Grau defined a translation from RCC8 into DL. This is achieved by defining a DL concept for every region and a set of translation rules for every RCC8 constructor [19].

Smart *et al.* and Abdelmoty *et al.* [26, 1] address the need for rules to extend geo-ontologies and facilitate spatial reasoning. This is mainly due to the limited expressivity of OWL. They identify two possible extensions, namely the formulation of spatial integrity constraints and rules for spatial relationships between objects in space. They developed a geo-ontology framework which is split into three components: a geo-ontology management system, a spatial reasoning engine, and an error management system [26, 1]. In [28] the authors describe a framework which focuses on four different ideas. The first approach deals with compiling the spatial relationships to the ABox and using nRQL as a query language. Another approach is a hybrid concept, where the ABox is associated with a *Space Box* (SBox) containing a set of spatial ground atoms which represent the whole spatial information. For querying the SBox, nRQL [16] is used, which is extended with *spatial query atoms*. In another approach, the ABox is extended again with an SBox, but spatial assertions are computed by means of inspection methods and materialized on the fly. The last method uses a standard ABox and exploits qualitative spatial reasoning, which is usable through spatial query atoms. All approaches were incorporated in the DL reasoner RacerPro for the DLMAPS system. PelletSpatial is a proof-of-concept implementation of RCC8 with a DL reasoner, featuring consistency checking and query answering. The authors extended the DL reasoner with a hybrid RCC8 reasoning engine, which is based on a path-consistency algorithm and a RCC8

composition table [27]. Finally, in the work of Grütter *et al.* a web search is enhanced with DL and spatial reasoning based on RCC8. RCC8 is encoded in SWRL-rules to capture spatial dependencies between different administrative regions [15].

SWRL was one of the first proposals for combining rules and ontologies. The rule layer in SWRL was set on top of an OWL KB by allowing material implication of OWL expressions [17]. Heterogeneous loose coupled approaches keep the rule base and DL KB as separate, independent components. The knowledge exchange is managed by an interface between the components. As a prominent example Description Logic Programs (dl-programs) can be taken, which were introduced by [11] and combine DL and normal logic programs under stable model semantics. Later they were extended in [10] to well-founded semantics. The concept of plug-ins in dl-programs was further generalized to HEX-programs [12] and lead to the successful development of the dlvhex reasoner.[1] In heterogeneous tight integrated approaches the combining of rules and DL is based on the integration of their models, where each model should satisfy its domain and agree with the other model. *CARIN* [20] and $\mathcal{DL}$+log [25] represent this approach. Full integrated approaches do not have any separation between the two vocabularies, this could either be achieved by a bidirectional translation of the different vocabularies or by rewriting both vocabularies to an overlapping formalism. Description Logic Programs (DLP) [14] and Hybrid MKNF knowledge bases [22] can be counted to this approach.

In the wider scope of our interest are Semantic Web search engines. For example, the authors of [13] developed two prototypes of a search engine. In the proposed systems an additional annotation step is used together with a domain specific ontology, with this step semantics is added to the elements of a web page.

3 Expected Contributions

Our contribution will be partly on the formal level and partly will include practical aspects. Derived from the research questions, we identify the following objectives:

- Creating a rule-based framework for combining heterogeneous spatial data, ontological reasoning, and spatial reasoning with focus on modular design and efficiency. Furthermore, we will consider non-monotonic features such as exception handling.
- Qualitative spatial reasoning will be considered, first by inferring regions out of points, and second by defining spatial relations among the regions. The spatial relations could be expressed in the well-know calculus RCC8. Furthermore, we will investigate the qualitative attributes of the inferred regions.
- The data integration should consider the semantical annotation and linking of heterogeneous data, and suitable ontologies for OSM, OGD, and other sources are needed. We will evaluate whether a modular or a centralized approach is more appropriate.
- Finally, we will provide a proof-of-concept implementation, which will be evaluated for query answering on large data sets and benchmarked against existing tools like PelletSpatial or DLMAPS.

We recognize, that objectives are quite challenging, particularly to find a good trade-off between a modular design and efficiency.

[1] http://www.kr.tuwien.ac.at/research/systems/dlvhex/

4 Proposed Methods

To fulfill the objectives, we have to put our focus on two issues. First, we need *a formal representation of the different abstraction levels* of spatial data. Based on the representation, *an inference mechanism for constraint checking and query answering* can be defined. Second, we need to outline an *architecture*, which is built around a multi-tiered reasoning engine. A central part of the architecture will be a top-level geo-ontology, which acts as an central repository for linking the different spatial data sources. The data sources could be defined in their own specific ontologies, which could be linked to the top-level ontology. The data sources will mostly be points in the metric space, but also other geometrical objects as lines and polygons appear.

4.1 Query Rewriting and Reasoning

We distinguish two different models for spatial data, namely point-based and region-based spatial models. The related spatial logics are considered either with reasoning about topological relations (interpretation over topological spaces) or about distances over metric spaces. We will start with point-based spaces and extend them by transformation to region-based spaces. These transformations could be calculated by a convex hull or by Voronoi tessellation [4]. The transformations require first metric spaces, but in addition also topological spaces by describing the relations between regions in RCC8. We identify a two-tier design to achieve a good trade-off between modularity and scalability.

Tier 1. The first tier is concerned with DL reasoning extended with point- and range-based spatial queries. At this point, there is no qualitative spatial reasoning, just a transformation (similar to the first step of [19]) of points from the spatial model to ABox assertions. Furthermore, we already consider the later described semantical annotation of points, which links the spatial data to DL concepts and roles.

Tier 2. The second tier is responsible for advanced transformation like Voronoi tessellation and qualitative spatial reasoning like RCC8. We propose to use HEX-programs, which facilitate declarative meta-reasoning through higher-order atoms. HEX-programs are an interesting candidate, because external atoms (e.g., description logic (dl) atoms) offer a query interface to other external computation sources (e.g., DL reasoning). Following from the idea of dl-atoms, we could extend HEX-programs with spatial atoms, which encapsulate access to the first tier and enables qualitative reasoning capabilities. Furthermore, by using HEX-programs under the answer-set semantics, we will be able to perform default and closed-world reasoning, translating and manipulate reified assertions, exception handling, and search in the space of assertions [12].

Dealing with Inconsistencies. Contrary to [26], we omit an error management system and perform inconsistency checks as a preprocessing step. ABox inconsistency is checked for linked spatial assertions by simply using the underlying DL reasoner. But as mentioned in [26], we have to deal with topological, directional, and duplicate inconsistencies. Stocker *et al.* [27] propose to check topological inconsistencies by calculating an $n \times n$ matrix M, where n are the different regions (represented as polygons), and using the *path-consistency algorithm* on M to approximate consistency. Duplicate inconsistencies occur with spatial objects of different sources, which are *intentionally* the

same, but are not aligned by `owl:sameAs`. We use custom heuristics to determine the similarity between objects, where the objects names, directions, shapes, and locations are considered. We leave directional inconsistencies for further research activities.

4.2 Architecture

Regarding technical aspects of the framework, we propose an architecture consisting of the following four parts:

Geo-Ontology. The OWL2 profiles OWL2 QL [7] and OWL2 EL [21] are interesting candidates for representing our spatial ontology. There has been a considerable effort of developing efficient query answering algorithms over ontologies using an RDBMS for both DLs. We draw on already defined work with GeoOWL for modeling a suitable geo-ontology. For our needs, the feature concept of GeoOWL is too general, thus we adopt a more detailed categorization based on GeoNames and OSM. There has been a large community effort to categorize the geospatial information in OSM, so we can derive with some additions an ontology out of the categorization. It is open how to unify the various OGD sources under a common ontology. However, most of the considered OGD sources of the city of Vienna are covered by the existing OSM concepts.

Knowledge Base. Heterogeneous sources of spatial data like OSM, OGD, or even local food guides could be integrated and linked by a central DL KB, which is part of our first tier. The KB should be based on our geo-ontology and act as a repository for the different sources keeping the vocabulary needed for concept and role queries. In particular, the annotated information should be kept in the KB, so we fulfill the modularity criteria and do not alter the original data sources. For example, for querying restaurants, the type of cuisine and atmosphere will be kept in the KB, but the geospatial information will be kept in the spatial database. The different spatial data sources can be stored in database like PostGIS keeping their native projections. Hence we will use a projection function to convert points to a predefined reference coordinate system.

Reasoning Engine. The first tier of the engine will incorporate a DL reasoner and a proprietary implementation for spatial predicates like *Near* or *Along*. From the vocabulary of our geo-ontology in combination with spatial functions, a join of a spatial and DL query will be created. We will exploit the rewriting techniques of OWL2 QL by compiling TBox and query into a SQL statement, which can be evaluated by an RDBMS over the ABox [7]. The second tier of the engine will mainly be build around dlvhex, an implementation of HEX-programs. We need at least one plugin to a computational geometry engine, one plugin to the first tier, and one plugin, which evaluates RCC8 relations on existing or on-demand calculated regions. Depending on the external data sources, a further plugin for accessing directly an RDBMS might be desirable.

Annotating Engine. A priori, the spatial data are not linked to any DL KB. Hence this step is crucial for integration them and extending the vocabulary of the queries. The spatial data is *linked* to the DL KB concepts and roles by asserting spatial objects to the ABox. For the OSM data, a straightforward task is to find related concepts, because our geo-ontology is partly derived from the OSM categories. For other data sources, we have to develop domain specific heuristics to assign spatial objects to our categorization (e.g. restaurant which are named *pizzeria* belong to Italian restaurants).

5 Conclusion and Future Work

We have proposed a novel framework for combining heterogeneous spatial data, ontological reasoning, and spatial reasoning. Our focus will be mostly on modularity, scalability, and non-monotonic features as default reasoning. Query Answering will not just cover standard POIs searches, but queries based on qualitative spatial reasoning (e.g., RCC8), which considers the inference of regions and the calculation of properties and spatial relations among the regions. For the data integration we consider the semantical annotation and linking of OSM, OGD, and further data sources by a top-level geo-ontology, which acts as a central repository. The above finding should lead to a proof-of-concept implementation with HEX-programs, which should be integrated into a larger research prototype containing a routing services and POI exploring facilities.

The ability to formulate constraints and defaults is a common goal for a knowledge representation formalism and increases expressivity quite dramatically. Using the loose-coupling approach to combining rules with external computation sources helps by facilitating modularity and allowing to integrate different DL reasoners and computational geometry engines quite naturally. However, compared to a tight coupling approach, using external computation sources usually has a negative effect on efficiency, as the external sources could be intractable, and certain optimizations and structural dependencies are easier to detect in homogeneous KBs.

Our future work will be along two parallel paths. Along the theoretical path, we have to investigate query rewriting and reasoning. Particularly the combination of modular HEX-programs, qualitative spatial reasoning, and DL needs to be addressed. We will also investigate how well qualitative spatial reasoning can be expressed in a rule-based languages. Furthermore, we will refine our centralized geo-ontology and investigate on a more modular design. Along the practical path, we will develop several dlvhex plugins, which enables access to spatial data and computational geometry functions. Furthermore, we will consider an implementation, which is more geared towards tractable evaluation. Finally, our implementation will be tested for efficiency and modularity in an artificial test environment and further under real-life conditions in an e-government and public transit system. The implementation is part of a larger project concerned with e-government and public transit systems.

References

1. Abdelmoty, A., Smart, P., El-Geresy, B., Jones, C.: Supporting frameworks for the geospatial semantic web. In: SSTD'09, pp. 355–372. Springer (2009)
2. Apt, K.R. and Warren, D.S. and Truszczynski, M.: The Logic Programming Paradigm: A 25-Year Perspective. Springer New York (1999)
3. Bennett, B.: Spatial reasoning with propositional logics. In: KR'94. pp. 51–62 (1994)
4. de Berg, M., Cheong, O., van Kreveld, M., Overmars, M.: Computational geometry: Algorithms and Applications. Springer (2008)
5. Bishr, Y.A.: Geospatial semantic web: Applications. In: Encyclopedia of GIS, pp. 391–398. Springer (2008)
6. Brewka, G., Eiter, T., Truszczyński, M.: Answer set programming at a glance. Commun. ACM (2011), to appear

7. Calvanese, D., Giacomo, G.D., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R.: Ontologies and databases: The dl-lite approach. In: Reasoning Web'09. pp. 255–356. Springer (2009)
8. Chen, H., Perich, F., Finin, T.W., Joshi, A.: Soupa: Standard ontology for ubiquitous and pervasive applications. In: MobiQuitous'04. pp. 258–267 (2004)
9. DeMers, M.N.: Fundamentals of geographic information systems (4. ed.). Wiley (2008)
10. Eiter, T., Ianni, G., Lukasiewicz, T., Schindlauer, R.: Well-founded semantics for description logic programs in the semantic web. ACM Trans. Comput. Log. 12(2), 11 (2011)
11. Eiter, T., Ianni, G., Lukasiewicz, T., Schindlauer, R., Tompits, H.: Combining answer set programming with description logics for the semantic web. Artif. Intell. 172(12-13), 1495–1539 (2008)
12. Eiter, T., Ianni, G., Schindlauer, R., Tompits, H.: Effective integration of declarative rules with external evaluations for semantic web reasoning. In: ESWC'06. pp. 273-287 (2006)
13. Fazzinga, B., Gianforme, G., Gottlob, G., Lukasiewicz, T.: Semantic web search based on ontological conjunctive queries. In: FoIKS'10. pp. 153–172. Springer (2010)
14. Grosof, B.N., Horrocks, I., Volz, R., Decker, S.: Description logic programs: combining logic programs with description logic. In: WWW'03. pp. 48–57. ACM (2003)
15. Grütter, R., Scharrenbach, T., Waldvogel, B.: Vague spatio-thematic query processing: A qualitative approach to spatial closeness. Trans. GIS 14(2), 97–109 (2010)
16. Haarslev, V., Möller, R., Wessel, M.: Querying the semantic web with racer + nrql. In: ADL'04 (2004)
17. Horrocks, I., Patel-Schneider, P.F.: A proposal for an owl rules language. In: WWW'04. pp. 723–731 (2004)
18. Jones, C.B.: Geographical information systems and computer cartography. Prentice Hall (1997)
19. Katz, Y., Cuenca Grau, B.: Representing qualitative spatial information in owl-dl. In: OWLED'05 (2005)
20. Levy, A.Y., Rousset, M.C.: Combining horn rules and description logics in carin. Artif. Intell. 104(1-2), 165–209 (1998)
21. Lutz, C., Toman, D., Wolter, F.: Conjunctive query answering in the description logic el using a relational database system. In: IJCAI'09. pp. 2070–2075 (2009)
22. Motik, B., Rosati, R.: A faithful integration of description logics with logic programming. In: IJCAI'07. pp. 477–482 (2007)
23. Renz, J.: Qualitative spatial reasoning with topological information, LNCS vol. 2293. Springer (2002)
24. Rodríguez, M.A., Cruz, I.F., Egenhofer, M.J., Levashkin, S. (eds.): GeoSpatial semantics, GeoS'05. Springer (2005)
25. Rosati, R.: Dl+log: Tight integration of description logics and disjunctive datalog. In: KR'06. pp. 68–78 (2006)
26. Smart, P.D., Abdelmoty, A.I., El-Geresy, B.A., Jones, C.B.: A framework for combining rules and geo-ontologies. In: RR'07. pp. 133–147 (2007)
27. Stocker, M., Sirin, E.: Pelletspatial: A hybrid rcc-8 and rdf/owl reasoning and query engine. In: OWLED'09. Springer (2009)
28. Wessel, M., Möller, R.: Flexible software architectures for ontology-based information systems. J. Applied Logic 7(1), 75–99 (2009)