A Semantic Mediator for Handling Heterogeneity of Spatio-Temporal Environment Data

Ba-Huy Tran, Christine Plumejeaud-Perreau, Alain Bouju, Vincent Bretagnolle

A semantic mediator for handling heterogeneity of spatio-temporal environment data

Ba-Huy Tran[1], Christine Plumejeaud-Perreau[2], Alain Bouju[1], and Vincent Bretagnolle[3]

[1] L3i, Université de La Rochelle, La Rochelle, France
[2] LIENSs, U.M.R. CNRS 7266, Université de La Rochelle, La Rochelle, France
[3] CEBC, U.M.R CNRS 7372,Université de La Rochelle, La Rochelle, France

Abstract. This paper presents the "Environment and landscape geo-knowledge" project which aims to exploit heterogeneous data sources recorded at the Chizé environmental observatory since 1994. From a case study, we summarize the difficulties encountered by biologists and ecologists experts when maintaining and analyzing environmental data, essentially the spatial organization of the landscape, crop rotation, and wildlife data. We show how a framework which uses a spatio-temporal ontology as a semantic mediator can solve challenges related to the analysis and maintenance of these heterogeneous data.

Keywords: data integration, ecology, environment, spatio-temporal ontology

1 Introduction

In rural areas with the predominance of agricultural activities, the study of environmental issues such as biodiversity preservation, soil erosion by water and tillage, erosive runoff, water pollution, and gene fluxes may benefit from the long-term analysis of the crop mosaic resulting from farming practices. In fact, agricultural landscapes are primarily designed by farmer decisions dealing with the crop choice and crop allocation at the farm scale. The arrangement, shape, and nature of crops compose the spatial organization of the landscape which impacts ecological processes at various scales. This information can be relevant when studying links between socio-economic environment, agricultural practices, and subsequent spatial organization of the landscape.

Recognizing the benefits of the long-term observation of agricultural practices for research on environmental issues, the UMR Chizé has established an observatory for crop rotation on the "Plaine & Val de Sevre" workshop area. Since 1994, a Geographic Information System for the Environment (GIS-E) has been deployed in order to monitor the crop rotation of agricultural parcels.

This paper presents first the context of this interdisciplinary research around the GIS-E. In the next sections, a spatio-temporal ontology, as well as a new framework, are proposed in order to improve the performance of the previous one and to solve challenges in spatio-temporal data analysis. Finally, the conclusion summarizes the progress achieved with the system while highlighting our future work.

1.1 Spatio-temporal environment data

For over twenty years, several databases have been collected by AGRIPOP teams (CNRS Chizé). These data can be categorized as follows.

Land use database The spatial organization evolves throughout time because farmers occasionally change the land use and boundaries of their parcels. Since 1994, the land use and spatial organization of 19,000 agricultural parcels are recorded from the field each year and centralized in a database that is initially modeled based on the Space-Time Composite paradigm[17]. The paradigm introduced a small geometry, here called *microparcel*, which is obtained by the intersection of all parcels during an observation period. The geometry of any parcel can be rebuilt "on the fly" by unionizing all *microparcels* belonging to it. The database contains over 600,000 records managed by the PostgreSQL DBMS extended with the PostGIS plugin.

Biology database Meanwhile, wildlife data are collected in the field for several years by another AGRIPOP team of Chizé. The data, timely and dated, comes from researchers who report their observations on over 600 species, mostly birds and plants, via their mobile devices. For birds, the base describes the behavior of observed species, their nests, and their contexts such as vegetation height, date-time, location, and weather condition. Over 26,000 observations are also managed by the PostgreSQL DBMS with its PostGIS extension.
There also exist numerous sets of structured data about different species, often in spreadsheets or in MS Access databases. The data concerns observations of ground beetles and small beetles which are auxiliaries of the fields and very sensitive to the quality of the environment. These insects have been monitored for over 9 years.

1.2 The need for spatio-temporal analysis

With the available data, a significant number of analyses can be conducted. These analyses, described as follows, require queries accomplished with spatio-temporal reasoning.

1. The analysis can be used first to verify the collected data sets. On crop rotation, experts can describe a certain number of succession rules in order to eliminate or correct questionable values. For example, the unlikely crop succession like "Sunflower-Sunflower" or "Sunflower-Rapeseed", as well as the disappearance of wood in the workshop area can be detected and examined. Primarily, this type of analysis needs temporal relationships reasoning between intervals of recorded land use statements.

2. On another hand, territorial events, such as fusion, integration, scission, extraction, reallocation, and rectification[18], are desired to be pointed out. Analyzing these events allows discovering the correlation between land use decision and land fragmentation or aggregation in farm practice. These events can be detected through spatio-temporal reasoning based queries.
3. Finally, experts also wish to seek the correlation between species observations and the land use of parcels. They could concern such animals' preferences by type and form of crop rotation. Cross-database queries with spatio-temporal relationships reasoning are required to select observations occurring in intervals of recorded land use statements.

2 SPATIO-TEMPORAL ONTOLOGY

We wish to develop an ontology which acts as a mediator to resolve the heterogeneities between these different data sources. Ontologies help to structure the knowledge and to improve the understanding of concepts through making clear how entities are linked to each other[11]. By defining entities and their relations, ontologies are considered as a feasible solution of the semantic heterogeneity problem[22], thus become the heart of semantic data integration systems[5]. The ontology of time and ontology of fluent are considered for this development.

2.1 Ontology of time and ontology of fluent

OWL-Time[4][14], dedicated to the concepts and temporal relationships as defined in the theory of Allen[2] and formalized in OWL, is certainly the best candidate. The ontology is used first to describe the temporal content of Web pages and temporal properties of web services. It is recommended by the W3C for modeling temporal concepts due to its vocabulary for expressing topological relations between instants and intervals. However, the ontology of time alone is not sufficient to represent the evolution of objects. Therefore, an upper-level ontology, such as the ontology of fluent which is based on temporal ontologies is strictly necessary.

Traditional ontologies are synchronic, i.e. they refer to a single point in time, thus the temporal dimension must be incorporated in order to monitor the spatial and semantic evolution of objects. Indeed, philosophers have distinguished between two paradigms: endurantism and perdurantism to represent diachronic identities. Endurantism assumes that objects (referred to *enduring* or *continuant*) have three dimensions and are available in full at every moment of their lives. Thus, these objects do not have the temporal dimension. In contrast, perdurantist approach considers objects (called *occurrent* or *perdurant*) to have four dimensions. These objects have several time slices in their lives constituting the temporal dimension. This approach represents the various properties of an entity over time as fluents that are only validated during certain intervals

[4] http://www.w3.org/2006/time

or instants. Therefore, perdurantist approach enables richer representation of real-world phenomenon through its flexibility and expressiveness[1].

The two main languages of the Semantic Web, RDFS and OWL, allow only binary relations between individuals, as a result, the temporal relationships between objects are neglected. The *4D-fluent* approach[23] has been proposed to overcome this limitation. The authors introduced the *TimeSlice* class to represent temporal parts of the entity which is linked to the *TimeInterval* class, a class of the time domain. Each entity is associated with an instance of the *TimeSlice* by the *tsTimeSliceOf* object property. This latter is connected to an instance of the *TimeInterval* by the *tsTimeInterval* property.

Several approaches based on the *4D-fluent* have been introduced. tOWL[6] extends OWL with a temporal dimension in order to allow the representation of complex temporal aspects, such as process state transitions. SOWL[3] extends OWL-Time by enabling representation of static as well as of dynamic information. Recently, the *Continuum* model[12] allows tracking the identity of spatio-temporal entities through time. This model has been successfully applied in studies of the urban evolution[12] or decolonization process[13].

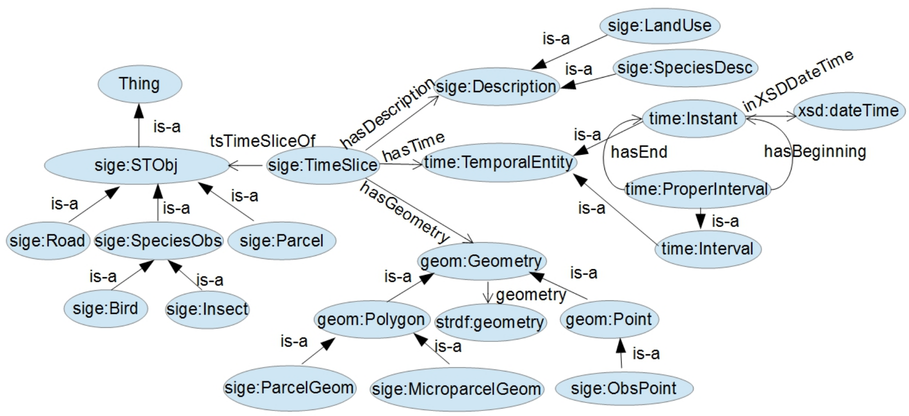


Fig. 1: A spatio-temporal ontology for environment

2.2 A spatio-temporal ontology for environment

We propose a spatio-temporal ontology (Fig.2) based on the *4D- fluent* approach that serves as a semantic mediator to integrate the presented datasets. This ontology is inspired by the *Continuum* model that examines the evolution of objects in both the temporal and spatial dimension. The main entities in our research, primarily parcels, roads and fauna and flora, have several time slices that match their different characteristics and spatial occupancies through their lives. In this way, crop rotation, boundary changes of each parcel, or species observations can be represented and analyzed.

While crop rotation or boundary changes of parcels are periodically archived by predetermined intervals of temporal, the position and behavior of species are collected at will. For this reason, the *4D-fluent* model is extended by generalizing the *Interval* class to the *TemporalEntity* class of OWL-Time that has two subclasses, *Interval* and *Instant*.

As presented, the land use database is built based on the *Space-Time Composite* paradigm which uses microparcel as a management unit. In consequence, we introduce the *MicroparcelGeometry* class as a subclass of the *Polygon* class that specializes the *Geometry* class. The difference in the spatial reference system used in these databases is an additional heterogeneity problem. Indeed, while the land use database uses the *NTF (Paris) / Lambert zone II* reference system for parcel geometries, observation points in the two other databases are recorded on the *WGS 84* one. This problem is handled in the mapping process which transforms the geometry data into virtual RDF triples and converts them into the same spatial reference system as well.

The following prefixes and associated URIs namespaces are used in the spatio-temporal ontology:

Prefix	**URI**	**Description**
sige	http://gemina.univ-lr.fr/owlSigE#	The spatio-temporal ontology for environment
geo	http://geovocab.org/spatial#	NeoGeo Spatial Ontology, a vocabulary for describing topological relations between features
strdf	http://strdf.di.uoa.gr/ontology#	The data model stRDF defining spatial datatype used in the Strabon triplestore
xsd	http://www.w3.org/2001/XMLSchema#	The W3C XML Schema Definition Language
time	http://www.w3.org/2006/time#	OWL-Time, an ontology for describing temporal concepts

2.3 Spatio-temporal reasoning

Qualitative relationships in the time domain are based on binary relations which are mutually exclusive. The work of Allen[2] introduced a temporal algebra to define topological relationships between dated objects. For two temporal intervals defined by their start and end date, there are the following relations: *before, meets, overlaps, during, starts, finishes* and their reverse, respectively *after, met-by, overlapped-by, contains, started by, finished-by*, and *equals* which does not have an inverse. These intervals can be viewed as instances of the *ProperInterval* class of OWL-Time. An interval is linked to two instants by the *hasBeginning* and *hasEnd* attribute that determine its boundaries. Besides these 13 relations, the *inside* relation between an instant and an interval must be also considered in order to link between databases.

To discover new temporal relations between objects, these above relations must be expressed by a set of rules. The *Semantic Web Rule Language* (SWRL[5]) is chosen due to its available libraries, called *built-ins*, that provide several predicates, mostly for date-time and duration processing. In this way, qualitative temporal relations between objects are derived by the Pellet[6] engine through a set of SWRL rules. This reasoning mechanism was applied in the SOWL[3] ontology which was afterwards improved by the CHRONOS[8] system. The SWRL rule corresponding to the *inside* relationship between an instant and an interval can be represented as follows:

```
Instant(?x), ProperInterval(?a), hasBeginning(?a,?b),
hasEnd(?a,?c), inXSDDateTime(?b,?d),inXSDDateTime(?c,?e),
inXSDDateTime(?x,?y), lessThanOrEqual(?y,?e),
greaterThanOrEqual(?y,?d)->inside(?x,?a)
```

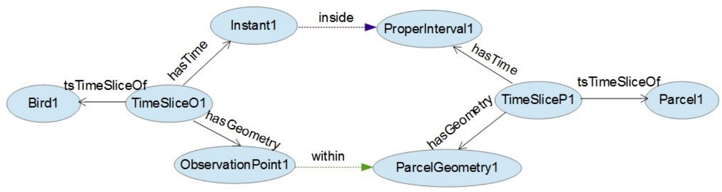
The spatial dimension of objects in our databases is represented by points and polygons which are defined by coordinates of points. In order to discover their spatial relations, qualitative relationships must be deducted from this quantitative information. In the literature, the topological analysis between spatial objects is often performed by the Nine-Intersection Model[7] or RCC8 model[19]. In both cases, we obtain an equivalent set of eight basic pairwise disjoint topological relations which are mutually exhaustive: *equals, disjoint,intersects, touches, within, contains* and *overlaps*.

Unfortunately, these relations cannot be inferred with simple SWRL rules. Several studies[15, 21] have introduced the SWRL *built-ins* for spatial relationships representation and processing, but there are still limitations with regard mainly to the system's performance and reuse capability. Therefore, in our project, the reasoning on complex spatial information is realized by a geospatial triplestore. Thus, spatio-temporal reasoning is accomplished through a combination of temporal SWRL rules and spatial functions of a triplestore.

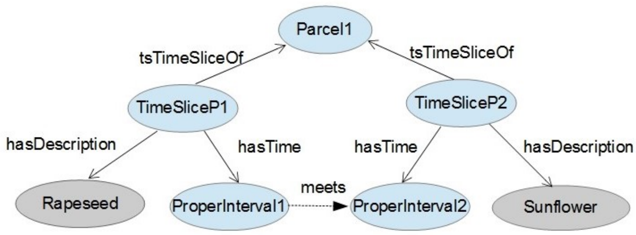
With the deducted spatio-temporal relations, the three major needs for data analysis can be fulfilled. Let's examine three simple corresponding cases below:

1. Species observations and land use data can be linked by combining the *inside* temporal relation between the instant of observations and the interval of recorded land use statements and the *within* spatial relation between observation points and parcel polygons (Fig. 2a).
2. Crop rotation can be verified by the *meets* temporal relation between intervals of land use statement of the same parcel (Fig. 2b).
3. Territorial events can be detected by incorporating the *meets* temporal relation between the interval of different timeslices and the *within* spatial relation between parcel geometries (Fig. 2c).

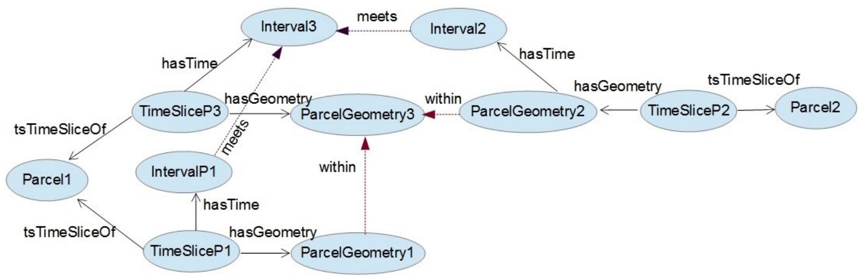
[5] http://www.w3.org/Submission/SWRL/
[6] http://clarkparsia.com/pellet/



(a) Linking of species observations and land use data.



(b) A 2-year crop rotation.



(c) An integration event between two parcels.

Fig. 2: Examples of data analysis.

3 A framework for spatio-temporal data analysis

In our previous work[20], a system architecture based on a translation technique was introduced. Nevertheless, the system cannot provide a promising response time since the translated queries do not exploit the strengths of the relational model nor the query optimizer. Furthermore, the selection conditions are not pushed down to the database[10]. As a consequence, triplestores are considered to improve the performance and functionality of the system.

Triplestores are DBMS for data modeled in RDF. Currently, several triplestores support storing and querying spatial data using GeoSPARQL or stSPARQL, extensions of SPARQL language. Those open-source that manage the best are uSeekM[7], Parliament[8] and Strabon[9][16]. Other triplestores support only a few type of geometries and geospatial functions[9]. Strabon is chosen since this open-source triplestore has a very good overall performance. This advantage can be explained by the push of the evaluation of SPARQL queries to the underlying spatially-enabled DBMS which has recently been enhanced with selectivity estimation capabilities[9]. Strabon extends the Sesame triplestore, allowing spatial RDF data stored in the Postgres DBMS enhanced with PostGIS. The triplestore works over the stRDF data model[16], a spatio-temporal extension of RDF in which the OGC standards, WKT and GML, are adopted to represent geospatial data.

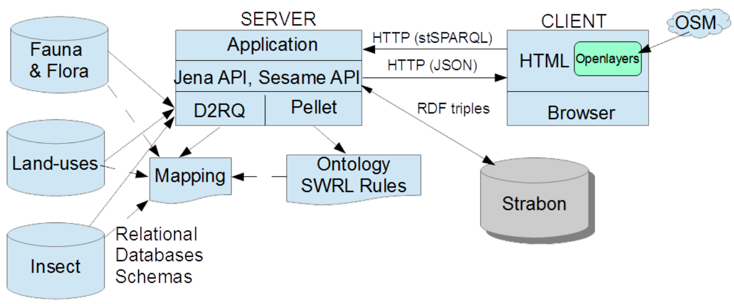


Fig. 3: A framework for environmental data analysis

A framework (Fig. 3) is developed, where a web server is hosted to receive stSPARQL queries from users in the form of HTTP requests. The framework consists of four parts: the data translation, temporal relation inference, triplestore bulk load, and data preparation and visualization.

1. **Data translation**: In order to populate the ontology with existing data sources, we rely on the translation techniques that define a mapping between

[7] http://dev.opensahara.com/projects/useekm/
[8] http://parliament.semwebcentral.org/
[9] http://strabon.di.uoa.gr/

databases and ontologies. The D2RQ[10][4] framework is chosen due to its support of different DBMS. The framework transforms relational data into virtual read-only RDF graph through a mapping file which describes how to connect to databases and to match our ontology to the databases schema. This RDF graph is then managed by the *Jena*[11] framework.

2. **Temporal relation inference**: The Pellet reasoner is used to deduce temporal relations between entities through a set of SWRL rules.
3. **Triplestore bulk load**: RDF triples are then imported to the Strabon triplestore that also hosts a SPARQL Endpoint.
4. **Data preparation and visualization**: RDF triples returned from Strabon are then prepared by *Jena* for visualization. The returned result is visualized through the *OpenLayers*[12] library with the geographical data from *OpenStreetMap*[13]. The results are stored in several different layers to facilitate the presentation and analysis.

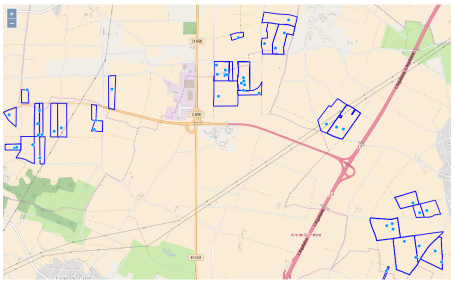


Fig. 4: A search for correlation between the nesting of Montagu's Harrier and different types of grassland in 2009.

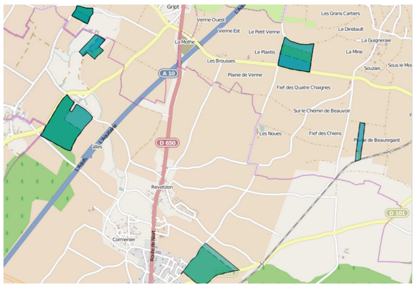


Fig. 5: A search for integration events of parcels in 2009.

```
select ?ts1 ?ts2 ?geom1 ?geom2
where
{
?p1 rdf:type sige:Parcel.
?p2 rdf:type sige:Parcel.
?ts1 sige:tsTimeSliceOf ?p1.
?ts2 sige:tsTimeSliceOf ?p2.
?ts1 sige:hasGeometry ?geom1.
?ts2 sige:hasGeometry ?geom2.
?ts1 sige:hasTime ?t1.
?ts2 sige:hasTime ?t2.
?t1 time:intervalMeets ?t2.
?ts1 sige:hasLandUse ?lu1.
?ts2 sige:hasLandUse ?lu2.
?lu1 sige:name "Rapeseed".
?lu2 sige:name "Sunflower".
}
```

4 Data analysis

The proposed framework along with the use of a spatio-temporal ontology as a semantic mediator can fulfill the three major needs of spatio-temporal analy-

[10] http://d2rq.org/
[11] http://jena.apache.org/
[12] http://openlayers.org/
[13] http://www.openstreetmap.org

sis. Indeed, the data model in the form of subjacent RDF graph facilitates the integration of different data sources. In addition, thanks to the Strabon triplestore and the Pellet engine, spatio-temporal relations between objects can be deducted. At the first time, only the land use and wildlife data are selected for experiments.

1. To analyze the correlations between crop rotation and biodiversity, experts can visualize the references of animals by type and form of crop rotation. For example, they can check out the correlation between the nesting of Montagu's harrier (*Circus pygargus*) and different type of grassland(Fig.4).
2. Through qualitative temporal relations inferred by the Pellet engine, researchers can also verify the quality of their data. Indeed, domain rules or expert knowledge on the crop rotation, appearance or disappearance of certain crop plants can be represented by stSPARQL queries to detect anomalies in collected data. For example, parcels having the hardly occurred succession "Rapeseed-Sunflower" can be located by the (Query 1).
3. Territorial events applied on farmland can be discovered by combining spatio-temporal relations. For example, integration events, in which a parcel has been absorbed by another, in 2009, can be retrieved and displayed on the map like (Fig.5). Since real parcel geometries can not be recorded with an absolute precision, the spatial relations between them can be converted to a more complex combination of other spatial relations and functions. In the latter example, the *within* relation is replaced by the *intersects* relation and the *area* and the *intersection* function.

These experiments are carried out on a 4 cores personal machines running at 2.8GHz with 8GB RAM. The performance of the new system is noticeably improved compared to the previous one[20]. The response time of a query for crop rotation decreases from 25 minutes to 5 seconds, thanks to the Strabon triplestore. Furthermore, the system supports now spatial reasoning through the stSPARQL language.

5 Conclusion

The presented work is part of the "Environment and landscape geo-knowledge" interdisciplinary project which sets out to improve the use of collected environment datasets on the "Plaine & Val de Sevre" workshop observatory since 1994. We seek to develop an open-source framework to exploit environmental data through semantic web technologies. We present an ontology and a framework that can fulfill the need for spatio-temporal analysis of these heterogeneous data. The proposed approach could be reused to perform management and analysis of long-term environmental data for other observatories.

In our perspectives, we consider integrating other datasets of the workshop area, such as insects and botanical data, or the satellite data. It will be then possible to use the system to enrich and qualify our data sources. We also plan to publish a portion of these data over the web as Linked Data in order to facilitate interchanges with other available datasets, especially the weather and infrastructure data concerning the workshop area.

References

1. Al-Debei M. M., Asswad M. M. al, Cesare S. de, Lycett M. (2012). Conceptual modelling and the quality of ontologies: Endurantism vs. perdurantism. CoRR, 2012.
2. Allen J. F. (1983). Maintaining knowledge about temporal intervals. Commun. ACM, Vol. 26, No. 11, p. 832843, 1983.
3. Sotiris Batsakis, Euripides G. M. Petrakis. SOWL: A Framework for Handling Spatio-temporal Information in OWL 2.0. Proceedings of the 5th International Conference on Rule-based Reasoning, Programming, and Applications, RuleML'2011, 2011.
4. Bizer C. D2rq - treating non-rdf databases as virtual rdf graphs. In In proceedings of the 3rd international semantic web conference (iswc2004), 2004.
5. Cruz I.F , Xiao H. The Role of Ontologies in Data Integration . Journal of Engineering Intelligent Systems, Vol 13, p. 245-252, 2005.
6. Frasincar F., Milea V., Kaymak U. towl: Integrating time in owl. In R. de Virgilio, F. Giunchiglia, L. Tanca (Eds.), Semantic web information management, p. 225-246, 2010. Springer Berlin Heidelberg.
7. Egenhofer M. J. and Herring J. R. Categorizing binary topological relations between regions, lines, and points in geographic databases. Technical report, Department of Surveying Engineering, University of Maine, 1990.
8. Eleftherios Anagnostopoulos and Sotiris Batsakis and Euripides G.M. Petrakis. CHRONOS: A Reasoning Engine for Qualitative Temporal Information in OWL.Procedia Computer Science, Volume 22, 2013, Pages 70-77, ISSN 1877-0509.
9. Garbis G., Kyzirakos K., Koubarakis M. Geographica: A benchmark for geospatial rdf stores. CoRR, Vol. abs/1305.5653, 2013.
10. Gray A. J., Gray N., Ounis I. Can rdb2rdf tools feasibily expose large science archives for data integration? In Proceedings of the 6th european semantic web conference on the semantic web: Research and applications, pp. 491505. Berlin, Heidelberg, SpringerVerlag, 2009.
11. Gruber T. R. A translation approach to portable ontology specifications. Knowledge Acquisition, Vol. 5, No. 2, p. 199 - 220, 1993.
12. Harbelot B., Arenas H., Cruz C. Continuum: A spatio-temporal data model to represent and qualify filiation relationships. In Proceedings of the 4th acm sigspatial international workshop on geostreaming, p. 7685. ACM, 2013.
13. Harbelot B., Arenas H., Cruz C. Un modle smantique spatio-temporel pour capturer la dynamique des environnements". 14 me confrence Extraction et Gestion des Connaissances, Rennes, France, 2014.
14. Hobbs J. R., Pan F. An ontology of time for the semantic web. ACM Transactions on Asian Language Information Processing, Vol. 3, p. 6685, 2004.
15. Karmacharya A., Cruz C., Boochs F., Marzani F. Use of geospatial analyses for semantic reasoning. In R. Setchi, I. Jordanov, R. Howlett, L. Jain (Eds.), Knowledge-based and intelligent information and engineering systems, Vol. 6276, p. 576-586. Springer Berlin Heidelberg, 2010.
16. Kyzirakos K., Karpathiotakis M., Koubarakis M. Strabon: A Semantic Geospatial DBMS, The Semantic Web ISWC 2012, p. 295-311. Springer Berlin Heidelberg, 2012.

17. Langran G. E., Chrisman N. R. A framework for temporal geographic information. Cartographica: The International Journal for Geographic Information and Geovisualization, Vol. 25, No. 3, p. 1-14, 1998.
18. Plumejeaud C. , Mathian H., Gensel J., Grasland C. Spatio-temporal analysis of territorial changes from a multi-scale perspective. International Journal of Geographical Information Science 25, 10 1597-1612, 2011.
19. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. In: KR (1992)
20. Tran B.H., Plumejeaud C.P., Bouju A., Bretagnolle V. Conception d'un système d'information géographique résilient pour l'environnement. Conférence internationale de Géomatique et Analyse Spatiale SAGEO 2014.
21. Vandecasteele A. Spatial ontologies for detecting abnormal maritime behaviour. In: OCEANS 2012 MTS/IEEE Yeosu Conference: The Living Ocean and Coast - Diversity of Resources and Sustainable Activities, Yeosu, South Korea, 2012.
22. Wache H. et al. Ontology-Based Integration of Information - A Survey of Existing Approaches. IJCAI-01 Workshop: Ontologies and Information, p. 108-117, 2001.
23. Welty C., Fikes R. A reusable ontology for fluents in owl. In Proceedings of the conference on formal ontology in information systems, p. 226-236. IOS Press, 2006.