BAKILLAH MOHAMED

REAL TIME SEMANTIC INTEROPERABILITY IN AD HOC NETWORKS OF GEOSPATIAL DATABASES: DISASTER MANAGEMENT CONTEXT

Thèse présentée
à la Faculté des études supérieures et postdoctorales de l'Université Laval, Québec
dans le cadre du programme de doctorat en Sciences géomatiques
pour l'obtention du grade de Philosophiae Doctor (Ph.D.)

DÉPARTEMENT DES SCIENCES GÉOMATIQUES
FACULTÉ DE FORESTERIE, DE GÉOGRAPHIE ET DE GÉOMATIQUE
UNIVERSITÉ LAVAL
QUÉBEC

2012

Résumé

Avec le développement rapide des technologies permettant la collecte et l'échange des données géospatiales, la quantité de bases de données géospatiales disponibles est en constante augmentation. Ces bases de données géospatiales représentent souvent une même réalité géographique de façon différente, et sont, par conséquent, sémantiquement hétérogènes. Afin que les utilisateurs de différentes bases de données puissent échanger des données et collaborer dans un but commun, ils doivent avoir une compréhension commune de la signification des données échangées et résoudre ces hétérogénéités, c'est-à-dire que l'interopérabilité sémantique doit être assurée. Il existe actuellement plusieurs approches visant à établir l'interopérabilité sémantique. Cependant, l'arrivée puis la récente prolifération des réseaux ad hoc modifient et rendent plus complexe la résolution du problème de l'interopérabilité sémantique. Les réseaux ad hoc de bases de données géospatiales sont des réseaux qui peuvent s'auto-organiser, pour des besoins ponctuels, sans qu'une structure particulière soit définie a priori. En raison de leur dynamicité, de l'autonomie des sources qui les composent, et du grand volume de sources disponibles, les approches dites « traditionnelles » qui ont été conçues pour établir l'interopérabilité sémantique entre deux sources ou un nombre limité et statique de sources ne sont plus adaptées. Néanmoins, bien que les caractéristiques d'une approche pour l'interopérabilité sémantique dans les réseaux ad hoc doivent permettre d'agir sur un grand volume de sources, il demeure essentiel de prendre en compte, dans la représentation de la sémantique des données, les caractéristiques particulières, les contextes et les dimensions spatiales, thématiques et temporelles des données géospatiales.

Dans cette thèse, une nouvelle approche pour l'interopérabilité sémantique en temps réel dans les réseaux ad hoc de bases de données géospatiales est proposée afin de répondre à la fois aux problématiques engendrées par les réseaux ad hoc et les bases de données géospatiales. Les contributions de cette approche pour l'interopérabilité sémantique en temps réel concernent majoritairement la collaboration dynamique entre les utilisateurs de bases de données géospatiales du réseau ad hoc, la représentation et l'extraction des connaissances, le mapping sémantique automatisé, la similarité sémantique et la propagation des requêtes dans le réseau ad hoc. Le cadre conceptuel qui sous-tend l'approche se base sur les principes de la communication dans les réseaux sociaux. À la suite du cadre conceptuel qui établit les fondements de l'approche, cette thèse présente un nouveau modèle de représentation des coalitions de bases de données géospatiales, dans le but de faciliter, dans un contexte d'interopérabilité sémantique, la collaboration entre les utilisateurs de bases de données géospatiales du réseau. Sur la base de ce modèle, une approche de découverte des sources pertinentes et de formation des coalitions se basant sur les principes de l'analyse des réseaux est proposée. Afin de gérer les changements du réseau en temps réel, des opérateurs de gestion des changements dans les coalitions sont proposés. Une fois les coalitions établies, les échanges de données entre les membres d'une même coalition ou de coalitions différentes ne peuvent être assurées que si la représentation de la sémantique est suffisamment riche et que les ontologies qui décrivent les différentes bases de données sont sémantiquement réconciliées. Pour ce faire, nous avons développé dans cette thèse un nouveau modèle de représentation des concepts, le soit le Concept multi-vues augmenté (Multi-View Augmented Concept - MVAC) dont le rôle est d'enrichir les concepts des ontologies avec leurs différentes contextes, la sémantique de leurs propriétés spatiotemporelles, ainsi que les dépendances entre leurs caractéristiques. An Ensuite, une méthode pour générer les concepts MVAC est développée, laquelle comprend une méthode pour l'extraction des différentes vues d'un concept qui sont valides dans différents contextes, puis une méthode d'augmentation du concept qui extrait les dépendances implicites au moyen d'algorithme de fouille de règles d'association. Ensuite, deux modèles complémentaires furent développés pour résoudre les hétérogénéités sémantiques entre les concepts MVAC. Dans un premier lieu, un modèle graduel de mapping sémantique automatisé, le G-MAP, établit les relations sémantiques qualitatives entre les concepts MVAC au moyen de moteurs de raisonnement basé sur des règles d'inférence qui intègrent de nouveaux critères de matching. Ce modèle se distingue par sa capacité à prendre en compte une représentation plus riche et complexe des concepts. Puis, nous avons développé un nouveau modèle de similarité sémantique, Sim-Net, adapté aux réseaux ad hoc et basé sur le langage de la logique descriptive. La contribution des deux modèles permet une interprétation optimale par l'utilisateur de la signification des relations entre les concepts de différentes bases de données géospatiales, améliorant ainsi l’interopérabilité sémantique. La dernière composante est une approche multi-stratégies de propagation des requêtes dans le réseau ad hoc, où les stratégies, formalisées à l’aide du langage Lightweight Coordination Calculus (LCC) qui supporte les interactions basées sur des normes sociales et des contraintes dans un système distribué, représentent différents moyens employés pour communiquer dans les réseaux sociaux. L’approche de propagation intègre un algorithme d’adaptation en temps réel des stratégies aux changements qui modifient le réseau.

L’approche a été implémentée sous forme de prototype utilisant la plateforme Java JXTA qui simule les interactions dynamiques entre des pairs et des groupes de pairs (réseau peer-to-peer). L’utilité, la faisabilité et les avantages de l’approche sont démontrés par un scénario de gestion de désastre naturel. Cette thèse apporte aussi une contribution supplémentaire en développant le nouveau concept de qualité de l’interopérabilité sémantique ainsi qu’un cadre pour l’évaluation de la qualité de l’interopérabilité sémantique en tant que processus. Ce cadre est utilisé à des fins d’évaluation pour valider l’approche. Ce concept de qualité de l’interopérabilité sémantique ouvre de nombreuses perspectives de recherches futures concernant la qualité des échanges de données dans un réseau et son effet sur la prise de décision.

Abstract

The recent technological advances regarding the gathering and the sharing of geospatial data have made available important volume of geospatial data to potential users. Geospatial databases often represent the same geographical features but from different perspectives, and therefore, they are semantically heterogeneous. In order to support geospatial data sharing and collaboration between users of geospatial databases to achieve common goals, semantic heterogeneities must be resolved and users must have a shared understanding of the data being exchanged. That is, semantic interoperability of geospatial data must be achieved. At this time, numerous semantic interoperability approaches exist. However, the recent arrival and growing popularity of ad hoc networks has made the semantic interoperability problem more complex. Ad hoc networks of geospatial databases are network that self-organize for punctual needs and that do not rely on any predetermined structure. "Traditional" semantic interoperability approaches that were designed for two sources or for a limited static number of known sources are not suitable for ad hoc networks, which are dynamic and composed of a large number of autonomous sources. Nevertheless, while a semantic interoperability approach designed for ad hoc network should be scalable, it is essential to consider, when describing semantics of data, the particularities, the different contexts and the thematic, spatial and temporal aspects of geospatial data.

In this thesis, a new approach for real time semantic interoperability in ad hoc network of geospatial databases that address the requirements posed by both geospatial databases and ad hoc networks is proposed. The main contributions of this approach for real time semantic interoperability are related to the dynamic collaboration among user agents of different geospatial databases, knowledge representation and extraction, automatic semantic mapping and semantic similarity, and query propagation in ad hoc network based on multi-agent theory. The conceptual framework that sets the foundation of the approach is based on principles of communication between agents in social network. Following the conceptual framework, this thesis proposes a new model for representing coalitions of geospatial databases that aim at supporting the collaboration among user agents of different geospatial databases of the network, in a semantic interoperability context. Based on that model, a new approach for discovering relevant sources and coalitions mining based on network analysis techniques is proposed. Operators for the management of events affecting coalitions are defined to manage real times changes occurring in the ad hoc network. Once coalitions are established, data exchanges inside a coalition or between different coalitions are possible only if the representation of semantics of rich enough, and the semantic reconciliation is achieved between ontologies describing the different geospatial databases. To achieve this goal, in this thesis we have defined a new representation model for concepts, the Multi-View Augmented Concept (MVAC). The role of this model is to enrich concepts of ontologies with their various contexts, the semantics of their spatiotemporal properties, and the dependencies between their features. A method to generate MVAC concept was developed. This method includes a method for the extraction of the different views of a concept that correspond to the different contexts, and an augmentation method based on association rule mining to extract dependencies between features. Then, two complementary models to resolve semantic heterogeneity between MVAC concepts were developed. First, a gradual automated semantic mapping model, the G-MAP, discovers qualitative semantic relations between MVAC concepts using rule-based reasoning engines that integrate new matching criteria. The ability of this model to take as input a rich and complex representation of concepts constitutes the contribution of this model with respect to existing ones. Second, we have developed Sim-Net, a Description Logic- based semantic similarity model adapted to ad hoc networks. The combination of both models supports an optimal interpretation by the user of the meaning of relations between concepts of different geospatial databases, improving semantic interoperability. The last component is a multi-strategy query propagation approach for ad hoc network. Strategies are formalized with the Lightweight Coordination Calculus (LCC), which support interactions between agents based on social norms and constraints in a distributed system, and they represent the different strategies employed to communicate in social networks. An algorithm for the real time adaptation of strategies to changes affecting the network is proposed.

The approach was implemented with a prototype using the Java JXTA platform that simulates dynamic interaction between peers and groups of peers. The advantages, the usefulness and the feasibility of the approach were demonstrated with a disaster management scenario. An additional contribution is made in this thesis with the development of the new notion of semantic interoperability quality, and a framework to assess semantic interoperability quality. This framework was used to validate the approach. This new concept of semantic interoperability quality opens many new research perspectives with respect to the quality of data exchanges in network and its impact on decision-making.

Remerciements

Merci tout d'abord à Dieu le miséricordieux

Avant de rentrer dans le vif du sujet concernant le sujet de l'interopérabilité sémantique dans un réseau ad hoc de sources de données ou de connaissances, il est temps pour moi de remercier les membres de mon réseau de connaissances. Je souhaiterais tout d'abord remercier mon directeur de thèse, le Dr. ing. Mir Abolfazl Mostafavi, professeur et directeur du centre de recherche en sciences géomatiques de l'Université Laval, pour son encadrement, sa vision et sa clairvoyance en ce qui concerne la problématique et l'avancement de ma thèse. Je le remercie aussi pour tous les longs moments où il a été patient avec moi pour me guider et m'encourager.

Je souhaite aussi remercier mon codirecteur le Dr. Jean Brodeur qui, malgré son emploi du temps chargé, a toujours répondu présent, et je lui témoigne ma sincère reconnaissance pour ses conseils qui étaient bénéfiques tout au long du déroulement de ma thèse.

Un grand merci au Dr. Yvan Bédard, professeur en sciences géomatiques de l'université Laval, qui m'a fait l'honneur d'être mon conseiller; je le remercie infiniment pour ses conseils constructifs et enrichissants.

Je remercie tous les membres du CRG, tous ceux que je connais de près ou de loin. Merci au Dr. Geoffrey Edwards qui tout d'abord était mon professeur dans un cours de cognition et qui m'a marqué par son professionnalisme, merci encore car il a accepté de faire la prélecture de ma thèse, laquelle constitue une étape très importante.

Je remercie finalement le Dr. Robert Jeansoulin, ex-directeur de recherche en informatique au CNRS et présentement attaché scientifique auprès de l'ambassade de France à Washington, D.C., qui m'a fait l'honneur d'être examinateur externe.

Par ailleurs, je remercie aussi le Dr. Steve Liang de l'université de Calgary pour sa confiance et son intérêt pour ma recherche.

Je remercie les membres du département d'informatique de Défense Canada à Ottawa et plus particulièrement l'ingénieur François Lemieux pour ses longues et intéressantes discussions concernant l'interopérabilité.

Je remercie mes parents pour leur amour et leurs prières, sans eux rien ne serait possible. Merci à mes frères et sœurs, merci à Jessica, Annie, Evmarei pour votre soutien et votre confiance.

Finalement, et si c'était à refaire, … heureusement pas!!!

À toute ma famille

Table of contents

List of Figures

List of Tables

Acronyms

CSP	Communicating Sequential Processes
CSS	Calculus of Communicating Systems
DAML	DARPA Agent Modelling Language
DL	Description Logics
KR	Knowledge Representation
LCC	Lightweight Coordination Calculus
MAS	Multi-agents System
MVAC	Multi-view Augmented Concept
MCP	Mapping Conflict Predicate
OIL	Ontology Inference Layer
OWL	Ontology Web Language
RDF	Resource Description Framework
SDI	Spatial Data Infrastructures
SIP	Semantic Interoperability Process
SWEET	Semantic Web for Earth and Environmental Terminology

CHAPTER 1
Introduction

1.1 Research Context

The recent technological advances in geospatial data gathering have resulted in a growing number of geospatial data producers. Combined with the increased pervasiveness and availability of various kinds of networks and Internet, the final result is that very high volumes of geospatial data are made available to end-users. At the same time, geospatial data remains costly to produce and maintain, so sharing the existing geospatial data is often put forward as a solution instead of producing more data. From this principle, the concept of *geospatial data reusability* has emerged, and with it, the need to assess whether geospatial data that was produced for a specific need and in a given context, is suitable for a geospatial data user that may have different requirements and operates in a different context. For geospatial data sharing and reuse to be meaningful, the parties must be aware of the meaning of their exchanged data. That is, semantic interoperability among geospatial data must be ensured.

In the geospatial domain, a well-known definition of interoperability is given in ISO TC204, document N271: interoperability is *"the ability of systems to provide services to and accept services from other systems and to use the services so exchanged to enable them to operate effectively together."* Semantic interoperability is also defined for systems as the "*knowledge-level interoperability that provides cooperating databases with the ability to resolve semantic heterogeneities arising from differences in the meaning and representation of concepts*" (Park and Ram 2004). It is also the analogous to communication and cooperation between humans (Brodeur et al. 2003; Kuhn 2005). Semantic interoperability is a fundamental building block of the Semantic Web, "*in which information is given well-defined meaning, better enabling computers and people to work in cooperation*" (Berners-Lee et al. 2001). Achieving semantic interoperability is mainly hampered by heterogeneity of geospatial data, which may be classified as syntactic (differences in formats), structural (differences in the structure of data) and semantic (differences in the intended meaning) (Stuckenschmidt 2003). The semantic heterogeneity problem has been mainly addressed by semantic mapping approaches, whose purpose is to discover semantic relations (also referred to as semantic bridges, semantic correspondences) between concepts describing data (Kalfoglou and Schorlemmer 2003; Euzenat and Shvaiko 2007). However, there is still no fit-all solution to this problem at the moment. But should it be successful, in this single user scenario, an isolated user should be able to discover among a large set of sources, the ones that fit best its requirements.

However, besides this well-recognized issue of heterogeneity, there is an emerging trend towards sharing geospatial data in the context of a collaboration, where various stakeholders (for example, government departments, local administrations, groups of citizens, scientists, etc.) share geospatial data and services to reach a common goal (such as producing risk maps of a given region). According to *social networking* principles, people and organizations need to collaborate in order to maximise the available resources and to reduce risks in decision-making processes that involve multiple parties (Fox 2008). Among typical entities that participate in collaboration involving geospatial data sharing, we find:

- Government and public organizations that use geospatial data for emergency, disaster, land use management, etc.
- Academics and researchers that study specific geographical phenomena;
- Private societies using geospatial data for marketing or other purposes;
- Individuals that use location-based services (for tourism, business management, transport, etc.) (Vaccari et al. 2009)

This has given rise to the concepts of *geocollaboration* and *collaborative geographic information systems* (*GIS*). Collaborative GIS encompass tools, theories, and methods that support participation of various stakeholders in the management of geographically distributed data (Balram and Dragicevic 2006). Disaster management is a relevant example to illustrate the need for ad hoc geocollaboration. Disaster management involves several phases, including risk assessment, prevention, planning, real time response to disaster and recovery phases. In disaster management, the organizations that have to collaborate are not necessarily known in advance, especially in the response and recovery phases. However, despite this fact, available resources provided by the different actors must be mobilized in an optimal manner. Communication and mutual understanding is critical in this context. For instance, geocollaboration can be established to produce risk maps, design evacuation plans in real time, localise hazards or secure zones, etc. Geospatial data are essential to accomplish those tasks.

In comparison with a simple, first scenario where a single user searches a set of sources for appropriate data, collaboration involves a higher degree of complexity, due to the semantic conflicts that arise among the stakeholders and refrain from establishing efficient collaboration. For example, two participants may define the geometry of a river as the surface covering the bed of the river, or as the overall surface covered by water. Therefore, achieving semantic interoperability for the purpose of collaboration involves resolving a larger range of conflicts than "simple" point-to-point semantic interoperability (Figure 1.1). In point-to-point semantic interoperability, only pairs of network members must align; agreements are reached only between two members. In a collaborative setting, one network member must take into account several other members at the same time, and a group must reach a common agreement which correspond to a common and often temporary requirement.

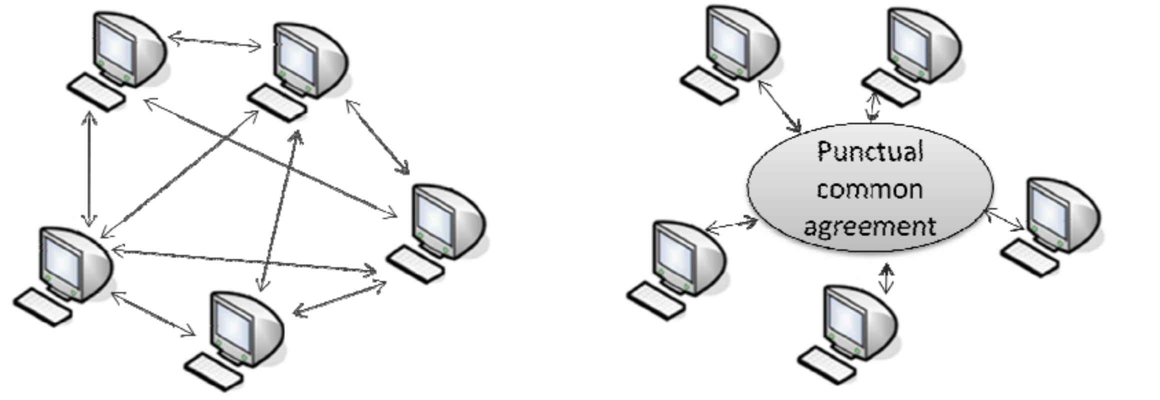


Figure 1.1a Point-to-point semantic interoperability

Figure 1.1b Semantic interoperability in a collaborative setting

Figure 1.1 Semantic interoperability between pairs of sources versus semantic interoperability in a collaborative setting

Ad hoc networks play an important role to support collaboration. The role of the network is to support data sharing by connecting members; its role is similar to the role of social networks. Advantages of ad hoc network are that new geospatial data sources can be added, while useless sources can be removed. Similar to ad hoc network are peer-to-peer networks where participants (called peers) have both capabilities of data and service consumer and provider, and may form groups (called super-peers) based on acquaintances. In recent years, popularity of ad hoc networks have significantly increased in various areas, such as e-commerce and e-government, and may be composed of different kinds of sources, including mobile devices and geo-sensor networks that monitor environmental phenomena (Nittel et al. 2004; Worboys and Duckham 2006). Semantic interoperability solutions that were developed for two sources or small sets of known sources are no longer suitable for ad hoc networks. Sources of an ad hoc network are autonomous and there is no a priori consensus among them, especially regarding how they describe their data (terminology used, level of detail, or granularity, domain of application, etc.). Consequently, the range of heterogeneity is wider. In addition, because of the dynamic nature of ad hoc networks, the additional issue of *real time* semantic interoperability must be addressed to support geospatial data sharing and collaboration in such environment. While a certain number of semantic interoperability approaches for networks and distributed environments have already been proposed (Hafsia 2001; Crespo and Gracia-Molina 2002; Löser et al. 2003; Staab et al. 2004; Zhuge et al. 2004; Zeinalipour-Yazti et al. 2005; Castano et al. 2006; Cudré-Mauroux 2006; Montanelli and Castano 2008), few are targeted at semantic interoperability in ad hoc networks of geospatial data sources.

1.2 Problem Statement

Semantic heterogeneity of geospatial data, which is the difference in the meaning of concepts representing geospatial data, is the main obstacle to semantic interoperability among geospatial databases. For example, concepts or their properties with different meanings may be given the same name, and different names may be used to designate concepts or properties representing the same reality. While syntactic approach have been put forward to support interoperability, and notably, the standards of the Open Geospatial

Consortium (OGC) for sharing data between different users based on common, widely-accepted formats and protocols, syntactic approaches do not address semantics and therefore are not sufficient to achieve semantic interoperability. There are two main issues that must be address to resolve semantic heterogeneity: knowledge representation and semantic reconciliation between different representations.

Knowledge representation is the issue of how to represent knowledge we have about a given reality, in an explicit and machine understandable way (Kavouras and Kokla 2008). In other words, it is the issue of how to represent semantics. Semantics is the "*meaning of expressions in a language*", or the meaning of symbols in a language (Kuhn 2005). The symbols of a language evoke concepts for users of this language, where concepts are abstractions of real world entities. The symbols in a language, in turn refer to the real world entities. The relations between symbols, real world entities and concepts are represented in the semiotic triangle (Ogden and Richard 1946). However, those relations depend on the context (Figure 1.2). For example, the symbol (term) "stream" may refer to a watercourse, or a flow of data, respectively in the context of hydrography and computer science.

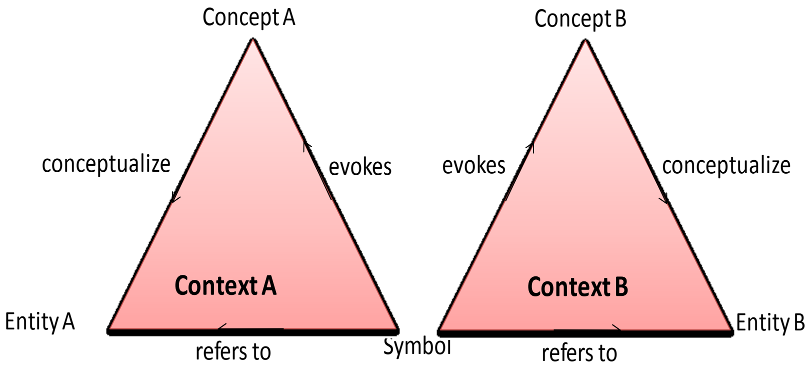


Figure 1.2 The relation between symbol, referee entity and concept depends on the context

Therefore, a symbol used by different users in different contexts may refer to different entities and evoke different concepts, creating semantic heterogeneity. Consequently, context is a fundamental element to be considered to resolve semantic heterogeneity (Brodeur et al. 2003).

Explicit semantics makes it possible to compare geospatial databases on a sound basis, while implicit semantics makes differences in meaning undetectable (Farrugia 2007) and therefore leads to misinterpretation of geospatial data contained in different databases. Ontologies, which are explicit specifications of a conceptualisation (Gruber 1993), have been proposed and adopted by many as a major component to support semantic interoperability because ontology capture the semantics of data in a machine understandable way (Brodeur et al. 2003; Kuhn 2003; Rodriguez and Egenhofer 2003; Agarwal 2005; Fonseca et al. 2005; Kavouras et al. 2005; Arpinar et al. 2006; Klien et a*l*. 2006; Lemmens 2006; Brodaric 2007). More concretely, an ontology provides a vocabulary that describes a domain of interest, called the universe of discourse, and specifies the meaning of terms with concepts, properties, relations, and axioms. Ontologies are now being considered as a main component of the Semantic Web, or Web 3.0. In this vision, the meaning of the data contained in websites would be explicitly represented and made available with ontologies, so that the Web would become a "Web of data" (Berners-Lee et al. 2001; Shadbolt et al. 2006). Ontologies can be used to describe the semantics of geo-services in order to discover relevant geo-services in a distributed environment (Lutz 2006). A common ontology can also be used to express common knowledge and reconcile different ontologies describing local resources (Fallahi et al. 2008). In the geospatial domain, ontologies are used to represent semantics of data (Brodaric et al. 2009). However, semantic heterogeneity of geospatial data is further complicated by the complex nature of geospatial concepts (Lemmens et al. 2006; Schwering 2008).

Furthermore, ontologies alone are not the complete solution, for they are themselves heterogeneous and also, their degree of semantic explicitness is varying (Obrst 2003). For example, ontologies include more or less rich semantic models, ranging from simple taxonomies to logic representations, which use for example First Order Logic (FOL) and Description Logics (DL). In addition, it is impossible and non-desirable, to produce an ontology on which everyone would agree (Kuhn 2005). In reality, members of society, which evolve in different social contexts, are able to communicate despite the fact that they do not have a formally-defined common language, because they have ways to resolve semantic ambiguities and differences in meaning. In computer science, semantic mapping approaches aim at fulfilling this task. Semantic mapping processes, which consists in discovering semantic relations between concepts of different ontologies or database schemas (Kalfoglou and Schorlemmer 2003), is considered as a plausible solution to resolve semantic heterogeneity (Vaccari et al. 2009). Most of the time, semantic mapping approaches aim to discover semantic relations based on set theory (equivalence, inclusion, overlap, etc.). Matchmaking is a reasoning process where the result can be a binary response (matching or not) or a similarity value (syntactic or semantic) (Kuhn 2005). Among semantic mapping and matchmaking approaches, very few consider spatial and temporal features of concepts in an explicit and separate manner. Therefore, they cannot detect and resolve spatial and temporal semantic heterogeneity, which is fundamental in the geospatial domain. In addition, many approaches use a simplified representation of concepts, not taking into account the complexity of geospatial concepts.

Beside the semantic heterogeneity problem, several other problems linked to the nature of ad hoc networks affect the semantic interoperability process. One of the expected advantages of ad hoc networks is that members can form groups to communicate, share geospatial data and collaborate to reach a common goal. The formation of groups has proven to be useful as it provides a way of structuring the network. The process of forming groups for the purpose of partitioning is called modularization. Modularization enhances the searching capabilities and the scalability of the searching process, as databases sharing similar content are already classified under a group's label. However, the purpose of groups is more than a way of partitioning the network; it also supports collaboration among network members (users of geospatial databases). In ad hoc networks, users of geospatial databases that can collaborate and exchange geospatial data do not known each other in advance, and therefore, the formation of groups is an issue that should be considered to achieve semantic interoperability. This issue includes several other issues, including defining such groups in theory, the representation of the semantics of groups, and techniques for discovering groups. The semantics of groups should also take into account theme, space and time.

One of the consequence of the dynamicity of ad hoc networks is that the solution for semantic interoperability must constantly adapt to changes in the network. Basically, changes to the ad hoc network are the addition or removal of a node (in this thesis, nodes are an abstract representation used for convenience). Concretely, addition of a node or its removal from the network may mean that:

- A geospatial database has been made available (or non-available) for security or privacy reasons, or because it has been updated (or it is outdated);
- A new geospatial database has registered to the network (or quit);
- In the context of ad hoc network composed of mobile devices, a node has entered (or left) the spatial neighbourhood of a group of nodes.

Following such changes, it is possible that existing groups must adapt by expanding, dividing, merging, etc. Strategies supporting changes in real time are required to ensure that queries sent through the ad hoc network by users are forwarded to relevant geospatial databases and that the required semantic mappings, which act as bridges between geospatial databases, are computed on-the-fly.

In summary, the general problem addressed in this thesis can be stated as follow:

The inadequacy of existing approaches for achieving real time semantic interoperability in ad hoc networks of geospatial databases.

This general problem can be decomposed in the following sub-problems, which will be addressed in this thesis:

- **The lack of a conceptual framework for real time semantic interoperability in ad hoc networks of geospatial databases.** While several frameworks for semantic interoperability exist, we are not aware of a conceptual framework that considers at the same time the specificities of geospatial databases, and the ad hoc network environment.
- **The lack of geospatial, semantic-based approaches for discovering and forming groups of geospatial databases, which take into account geospatial aspects of the**

coalitions. Several approaches have been proposed to support the formation of groups in various kinds of networks, mostly peer-to-peer networks (Khambatti et al. 2002; Agostini and Moro 2004; Bloehdorn et al. 2005; Kantere et al. 2008) or multi-agents systems (MAS), where the resulting groups are described as the results of a negotiation process (Zheng et al. 2008; Boella et al. 2009). Often, the semantic component of those groups is not considered or weakly represented.

- **The inadequacy of existing geospatial concept representation approaches, especially with respect to the representation of spatial and temporal semantics, views of the geospatial concept prevailing in various contexts and complexity of the concept's structure**. Geospatial concepts are more complex than concepts for thematic-only application domains. Geospatial concept definitions include spatial relations (for example, "floodplain is a meadow that is adjacent to a river") (Kavouras and Kokla 2008), geometry descriptions, in addition to temporal properties, which have their own semantics. The same geographic feature can be represented using different geometries (point, line, polygon, geometries with different dimensions, etc.), and furthermore, the same type of geometrical primitive used to represent a given object can represents different parts of this object. Other particularities of geospatial concepts include the fact that there are different ways to define a concept, depending on the context (Parent et al. 2006). For instance, roads may be represented with lines for transport planning purpose, but with polygons for road construction planning. While the context is a fundamental aspect for defining the meaning of geospatial concepts, it is seldom explicitly represented in geospatial databases or geospatial ontologies. In addition, several concept representation approaches assume that concepts are unstructured sets of features (Rodriguez and Egenhofer 2003; Schwering 2008) because properties are independent from each other. However, geospatial data is characterized by implicit linking (Lemmens et al. 2006), for example implicit dependencies between characteristics of concepts such as dependency between level of flooding risk and altitude.

- **The insufficiency of existing semantic mapping approaches with respect to the complexity of the underlying concept representation**. Many semantic mapping approaches use a simplified representation of the concepts that are compared, for example, with taxonomic relations only. Consequently, their matching criteria are insufficient to compare more complex geospatial concepts such as described above.

- **The absence of semantic similarity measure for ad hoc networks of geospatial databases**. In addition, to the best of our knowledge, the literature review indicates that there are no semantic similarity measures that were specifically dedicated to ad hoc networks. Such a semantic similarity measure would have to take into account the high semantic heterogeneity level of domains of applications covered by the ontologies to match, the possibility of exploiting neighbour ontologies in the ad hoc network, and a procedure to adapt the similarity to changes that occur in the ad hoc network (such as the addition or the removal of a database).

- **The need for different real time geospatial query propagation strategies that forward queries to relevant databases and adapt to changes in the ad hoc network**. Many existing semantic mapping approaches dedicated to dynamic networks focus on the issue of automating the discovery of semantic mappings. However, this is not the only issue. For scalability reasons, semantic mappings cannot be computed for all pairs of databases, even if the process is automatic. Therefore, a method for determining what databases have to be semantically reconciled is required. The need for computing semantic mappings arises when a geospatial query is formulated by the user of some geospatial database. Then, we must determine to which geospatial databases the query should be forwarded, in order to obtain optimal results. Once the query is propagated to the relevant sources, semantic mappings can be computed to find the relevant concepts in the ontologies of these sources. In fact, several query propagation approaches rely on semantic mappings to determine the propagation path; this type of approach does not resolve the problem of the large volume of semantic mappings that must be computed.

- **The absence of a framework defining the quality of semantic interoperability processes in ad hoc networks of geospatial databases**. Assessing the quality of the semantic interoperability process is an important issue that must be addressed in parallel with the development of semantic interoperability frameworks. A number of researches have been conducted on quality of geospatial data (Couclelis 2003; Devillers et al. 2005; Devillers et al. 2007; Goodchild 2007; Congalton and Green 2009; Sadiq and Duckam 2009) and quality of ontology (Cross and Pal 2005; Mostowfi and Fotouhi 2006). However, no framework was proposed to assess quality of semantic interoperability as a process. Therefore, the user that gathers data from various sources using semantic interoperability processes is unaware of the quality of this process, and therefore, of the suitability of retrieved data.

1.3 Hypotheses and Research Objectives

1.3.1 Hypotheses

The main hypothesis of this thesis the following: real time semantic interoperability in ad hoc networks of geospatial databases can be **improved** if (1) the ad hoc network is partitioned into meaningful coalitions of geospatial databases; (2) the ontologies of databases are semantically augmented by making explicit the implicit semantics; (3) a semantic reconciliation approach can process all features of the semantically augmented databases' ontologies; and (4) a semantic-aware query propagation approach can combine different strategies, which use different types of knowledge, to reach relevant databases of the network.

1.3.2 Objectives

To validate the hypotheses that were made, the general objective of this thesis is to propose a **framework and an approach for real time semantic interoperability in ad hoc networks of geospatial databases**. The specific objectives are:

- To propose a conceptual framework for real time semantic interoperability in ad hoc networks of geospatial databases. This objective is achieved in Chapter 3.

- To propose a semantic model for coalitions of geospatial databases; to develop, based on the proposed model, an approach for discovering relevant coalitions of geospatial databases in ad hoc networks. This objective is achieved in Chapter 4.

- To develop a geospatial concept representation approach that integrate elements cited in section 1.2, including semantics of spatiotemporal features, different contexts of a concept and takes into account the fact that the concept is a set of structured features; to develop a semantic augmentation method and tool that supports the generation of such concept. This objective is achieved in Chapter 5.

- To propose a real time query propagation approach that determines the relevant databases to which a given geospatial query must be submitted, and that established an order over these relevant databases; to develop an algorithm to adapt the query propagation to the real time changes that occur in the ad hoc network. This objective is achieved in Chapter 6.

- To propose a semantic reconciliation approach that includes (1) a qualitative semantic mapping approach and tool for ad hoc networks that is adapted to the concept representation developed in the previous specific objective, and (2) a semantic similarity measure for ad hoc networks of geospatial databases, that support the most prominent representation ontology language, Description Logics (DL). This objective is achieved in Chapter 3 and Chapter 7.

- To develop a conceptual framework for quality of semantic interoperability in ad hoc networks of geospatial databases, and quantitative measures to assess the quality of the semantic interoperability. However, the objective is not to develop a comprehensive set of measures for assessing quality of semantic interoperability, but rather a framework that proposes the basis of what is quality of semantic interoperability.

1.4 Overview of the Methodology

- ***Phase 1: Literature Review and Definition of the Research Problem***

  This phase was dedicated to the study of related research on semantic interoperability. First, we studied the notions related to semantic interoperability, including the proposed definitions of semantic interoperability and semantics, and the problems hampering semantic interoperability, including heterogeneities, and the dynamicity of the ad hoc networks. Then, we reviewed existing global frameworks for semantic interoperability, included those for limited and determinate number of sources, and those for networks. In addition, we studied the concept of geocollaboration and approaches for the discovery and the formation of groups in various types of networks; we noted that most of the work in this field is dedicated to peer-to-peer networks, while, to the best of our knowledge, very few approaches are dedicated to the formation of groups of geospatial databases. Other research areas that have been investigated include semantic representation approaches, and more specifically, existing concept representation approaches, as well as concept representation approaches underlying semantic similarity frameworks. Existing semantic similarity measures were also presented. Then, the large domain of semantic mapping frameworks, techniques, approaches and tools was thoroughly studied, starting with existing reviews by Kalfoglou and Schorlemmer in 2003 and Euzenat and Shvaiko in 2007, to investigate more recent work. It was concluded that existing semantic mapping criteria were insufficient to compare complex geospatial concepts. At last, techniques for propagation of queries in networks were investigated. It was determined that a combination of complementary real time geospatial query propagation techniques is required for ad hoc networks.

- ***Phase 2: Developing a Conceptual Framework for Real Time Semantic Interoperability in Ad Hoc Networks of Geospatial Databases***

  This phase is aiming to identify the required components for such a framework. Our methodology consists in using the principles of social networks as a theoretical foundation for our framework. The real time semantic interoperability process is seen as the communication between members of a social network. From characteristics of social networks, semantic interoperability principles were derived, which provided the necessary guidelines for the framework. The developed framework formalizes the required components, including an approach for representing coalitions of geospatial databases, a multi-view and augmented geospatial model for concepts that compose the ontologies which describe the semantics of data, and a semantic mapping model adapted to the proposed geospatial concept representation. In addition, the framework outlines the different strategies for propagating geospatial queries to relevant geospatial databases, which are based on strategies used by members of social networks to communicate and find resources.

- ***Phase 3: Presenting a Model for Geospatial Database Coalitions and Approach for Discovering Geospatial Database Coalitions in Ad Hoc Networks*.**

  As it was identified in the literature review, existing representations of groups of databases (or peers, or agents) give limited attention to representing semantics of such groups. In this third phase, we have developed a model of a coalition of geospatial databases, where the characteristics of coalitions related to theme, space and time are represented as constraints. The model is used as the semantic basis of a coalition mining algorithm that discovers the coalitions based on network analysis techniques. The algorithm reproduces the tendency of members of social networks to group around members having more leadership.

- ***Phase 4: Developing a Geospatial Concept Representation Approach and a Method for Generating Such Concepts***

  We have adopted a multi-view representation approach to represent the different contexts under which a geospatial concept can be considered. One of the intended purposes of representing different contexts of the concept was to allow the user to select the relevant context with respect to its expectations. Also, the multi-view representation approach introduces the idea of multi-context semantic interoperability. In addition, we proposed to integrate an augmented part of the concept which represents dependencies between features of the geospatial concept. The developed geospatial concept model was named the Multi-View Augmented Concept (MVAC) Model. A method called MVAC augmentation, which integrates data mining techniques and Semantic Web techniques, was developed to generate MVAC concepts from "original" concepts in ontologies.

- ***Phase 5: Development of Real Time Query Propagation Strategies in Ad Hoc Networks of Geospatial Databases***

  We have conceptualized and developed social-network-inspired real time strategies for propagating geospatial queries to relevant geospatial databases. The methodology for developing the strategies is to use the processes by which members of social networks communicate and disseminate information through the network, including using the knowledge content held by people, the historical memory of people, the knowledge that people have about people, and the organizational properties of society.

- ***Phase 6: Proposing a New Semantic Reconciliation Approach composed of a Semantic Mapping Approach and Semantic Similarity Model for Ad Hoc Networks of Geospatial Databases***

  Finally, to resolve semantic heterogeneity between geospatial databases of the ad hoc network, we need to develop an adapted semantic mapping approach that will have the capacity to compare MVAC concepts. In the literature review, it was determined that existing semantic mapping approaches are unable to handle this task. Consequently, new matching criteria which use the augmented part of the MVAC and the multiple views were developed and integrated into the semantic mapping model. The semantic mapping approach is named G-MAP, where G stands for the gradual process by which MVAC concepts are compared. G-MAP is based on a rule-based reasoning process inspired from inference engines. The G-MAP requires a quantitative counterpart that measures the level of similarity between concepts. In parallel with the development of G-MAP, a semantic similarity model for ad hoc networks of geospatial databases was also developed, named SIM-NET; it is based on Description Logics (DL) language.

- ***Phase 7: Implementation and validation of the approach with a Prototype and Experimentation with a New Framework for Quality of Semantic Interoperability***

  Finally, we have developed a prototype which implements and integrates the coalition discovering approach, the query propagation approach, the MVAC semantic augmentation approach and the semantic reconciliation approach to achieve real time semantic interoperability in ad hoc networks of geospatial databases.

The validation with the prototype and the experimentation consists of the following steps:

1) Implementing a Java tool for the discovering and creation of coalitions of geospatial databases.

2) Developing a Java implementation of the real time propagation strategies and testing their respective performance for random set of queries sent in the network. The objective was to compare the proposed strategies to determine their respective strengths and weaknesses.

3) Implementing the MVAC Java tool for the generation of MVAC concepts. The MVAC tool involved the implementation of a view extraction approach and a dependency extraction approach to semantically augment the concept.

4) Implementing the semantic reconciliation Java tool for computing semantic mappings and semantic similarity among MVAC concepts.

5) Integrating the tools and algorithms developed in steps 1 to 4 into a single prototype built on the JXTA platform, which is an open source Java platform that simulates an open and dynamic peer-to-peer network, where peers can form groups, send messages to other peers, enter or quit the network. The usefulness of the prototype was validated within a disaster management scenario.

6) Developing a framework for assessing the quality of semantic interoperability process;

7) Applying this framework to perform experimentation on the prototype and show that the proposed approach improves real time semantic interoperability in ad hoc networks of geospatial databases.

Figure 1.3 illustrates the steps of the methodology followed in this thesis.

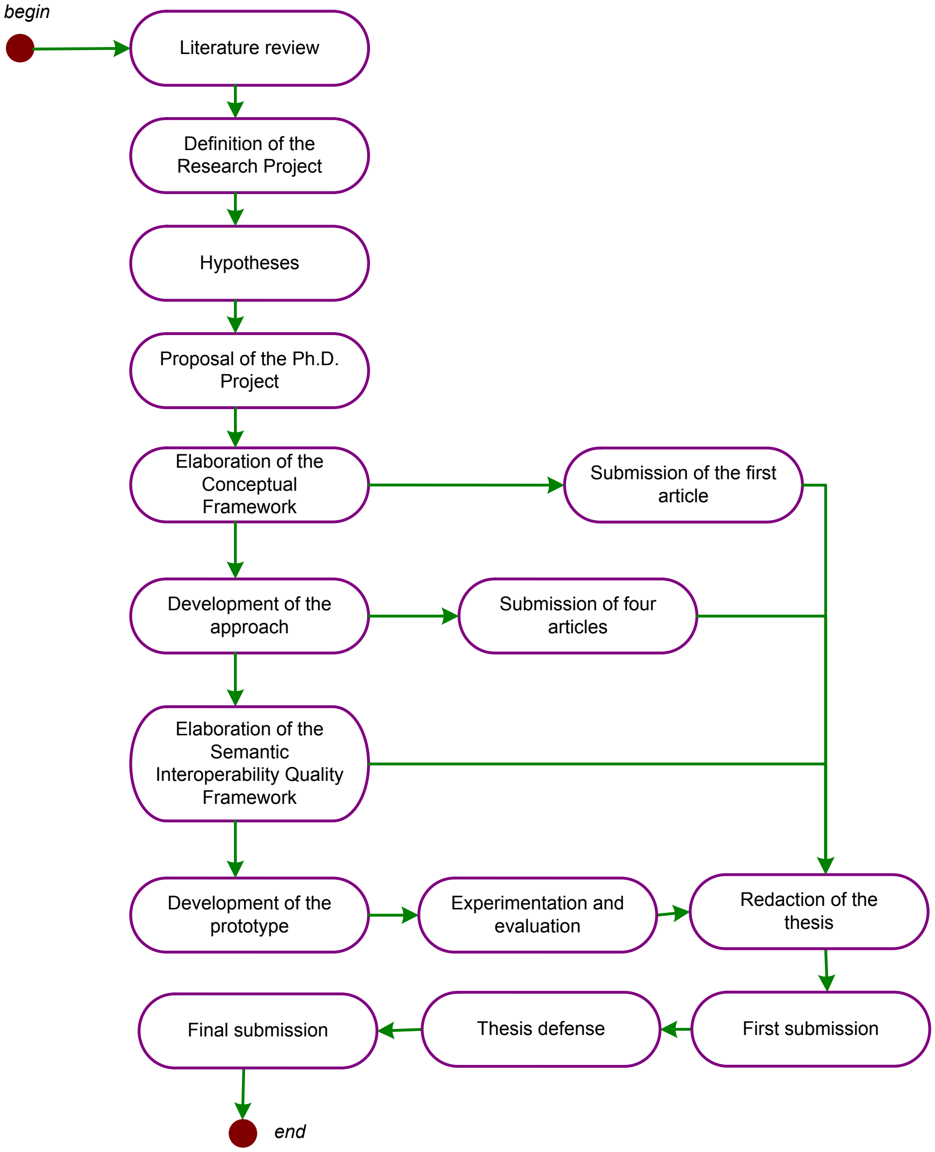

begin
Literature review
Definition of the Research Project
Hypotheses
Proposal of the Ph.D. Project
Elaboration of the Conceptual Framework
Submission of the first article
Development of the approach
Submission of four articles
Elaboration of the Semantic Interoperability Quality Framework
Development of the prototype
Experimentation and evaluation
Redaction of the thesis
First submission
Thesis defense
Final submission
end

Figure 1.3 Schema of the research methodology

1.5 Organisation of the Thesis

This thesis has resulted in 7 articles. Two articles that compose Chapters 5 and 7 were accepted and published in scientific journals; one article that composes Chapter 3 was published in a conference with peer review committee and one article that composes

Chapter 5 was published as a book chapter; three articles that compose Chapters 3, 4 and 6 were submitted to scientific journals.

The second chapter presents the background of the research, including a presentation of ad hoc networks and their characteristics, semantic-related subjects such as ontologies and semantic interoperability, a review of relevant data mining techniques used in this thesis, and a literature review on semantic interoperability frameworks, approaches and tools. The next chapters present the contributions of the thesis, which were published in several original papers. The third chapter presents the conceptual framework of the real time semantic interoperability approach. It includes the conceptual models for geospatial databases coalitions, the MVAC geospatial concept model, and the G-MAP augmented semantic mapping approach. The G-MAP was also validated as an application for the semantic interoperability of geospatial web services in SeCOGIS 2011, which was selected as best paper. This chapter has been submitted as an article in the *International Journal of Geographical Information Science* (IJGIS). The fourth chapter is a paper that presents the approach for discovering geospatial databases coalitions, including the coalition mining algorithm and operators for the management of dynamic coalitions. It has been submitted to the *Data and Knowledge Engineering* (DKE) journal. The fifth chapter is a paper that presents the MVAC tool extraction and generation tool which is based on the MVAC conceptual model presented in the conceptual framework. This chapter has been accepted and will be published in the *Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science*, Hong Kong, 26-28 May 2010, as well as selected to be part of the *ISPRS 2011* book. An extended version was accepted and will be published in the *Journal of Earth and Engineering* in 2012. The sixth chapter presents that framework, implementation and testing of the real time query propagation strategies, and was submitted to *Journal of Network and Computer Applications*. The seventh chapter is a paper that was published in 2009 in *Transactions in GIS* and presents SIM-NET, a semantic similarity approach for ad hoc network of geospatial databases. The eighth chapter presents the implementation of the prototype. The ninth chapter presents a framework for quality of semantic interoperability, which extends a framework for quality semantic mappings that we have published in the *International Symposium for Spatial Data Quality 2007* (ISSDQ

2007), and that was selected as one of the best papers to be published in *Quality Aspects in Spatial Data Mining* (Bakillah et al. 2009). Chapter 9 also presents further experimentation of the prototype with the semantic interoperability quality framework. The last chapter presents the conclusion of the thesis, including future research perspectives. The papers that were published and that compose the thesis have been very slightly modified after being integrated in the thesis. Consequently, the content of some chapters may seem redundant, but this is only to ensure that each article stands by its own and presents adequate background and context.

CHAPTER 2

Background

2.1 Résumé du chapitre

Ce chapitre présente les différentes notions qui sont à la base de la recherche présentée dans cette thèse. Ce chapitre présente également une revue de littérature des domaines dans lesquels des contributions ont été faites. Premièrement, les réseaux ad hoc, leurs caractéristiques et leurs rôles sont introduits. Dans la seconde section, nous présentons les notions liées à l'interopérabilité sémantique, notamment les problèmes qui empêchent de réaliser l'interopérabilité sémantique, ainsi que le concept d'ontologie, des languages et des raisonnements. Ensuite, une revue de littérature sur les domaines contribuant à l'interoperabilité sémantique est fournie, plus précisément sur la découverte et la formation de groupes dans un réseau, la représentation et l'extraction des connaissances, la similarité sémantique, le mapping sémantique et finalement, les techniques de propagation des requêtes dans un réseau.

2.2 Introduction

The research presented in this thesis is based on several notions that will be presented in this chapter. This chapter also presents a literature review on the domains for which a contribution was made in this thesis. First, we present the ad hoc networks, their characteristics and roles. The second section is a background on semantic interoperability, including problems related to semantic interoperability and ontologies. Then, a state-of-art on issues related to semantic interoperability is presented. In this literature review, we discuss the issues of group discovering and formation, knowledge representation and extraction, semantic similarity, semantic mappings and finally query propagation in networks.

2.3 Ad Hoc Networks

2.3.1 Ad Hoc Network Characteristics

Ad hoc networks represent a recent computing paradigm that enables the rapid, on-the-fly formation and dissolution of networks with short existence. An ad hoc network is formed by a mobile platform, which is composed by nodes that represent autonomous systems (Hafsia 2001). Ad hoc networks require no fixed infrastructure, and their nodes are self-organizing into temporary configurations for often short-term purposes. In this thesis, we consider autonomous nodes of the ad hoc network to be geospatial databases, which may be fed by various data-producing devices, including wireless mobile devices, sensors and geo-sensors, etc. The main characteristics of ad hoc networks, which may influence interoperability, are summarized as follow:

- **Nodes of the ad hoc networks are autonomous.** The autonomy of nodes can be understood from different points of view: from the perspective of mobility, nodes are autonomous because they are free to move, they can be available or unavailable at any time, and they can quit or enter the network at any time. From the perspective of content, nodes are autonomous because there is no a priori agreement between them regarding how they describe the data they hold, and how each of them represents the real world. In particular, this means that they make different ontological commitments[1].

- **Ad hoc networks have a dynamic topology.** This is due to the fact that nodes are free to move, and therefore, the members and configuration are not predictable. This means

[1] an ontological commitment is “an agreement to use a vocabulary ... in a way that is consistent ... with respect to the

- that semantic interoperability strategies constantly need to adapt to the current available nodes and topology.

- **Ad hoc networks may have bandwidth constraints.** For ad hoc network composed of wireless mobile devices, links have significant lower capacity than other types of network. This means that the transmission rate is low, which impact the amount of communications that can be made. This characteristic also impacts the semantic interoperability strategy, which must restrain the number of nodes being accessed to answer a given query.

- **Nodes of ad hoc networks are prone to security threats.** Nodes of the ad hoc network are necessarily more prone to physical security threats than fixed computer networks, so nodes may suddenly become unavailable without warning. The semantic interoperability strategy must therefore adapt in real time to such changes.

The major advantage of ad hoc networks, which is dynamicity, turns out to be a constraint that requires the development of adapted semantic interoperability solutions. Nevertheless, the role of ad hoc networks in geospatial data sharing is still fundamental.

2.3.2 Roles and Applications of Ad Hoc Networks in the Geospatial Domain

Fundamentally, the role of networks is to connect participants in a flexible manner in order to maximise the likeliness of discovering relevant resources. To enable this discovery of relevant resources, the role of a node in an ad hoc network may be to:

- act as an information requestor, i.e. sending queries to other databases of the network;
- act as an information provider, i.e. answering queries submitted by other nodes of the network;
- act as a relay in forwarding queries between a requestor node and a provider node.

Many geospatial applications require the dynamic and flexible communications capabilities provided by ad hoc networks. Ad hoc networks can be employed when there is no other available communication infrastructure to exchange geospatial data, such as following human or natural disasters. Ad hoc networks play a fundamental role in different kinds of applications requiring mobility, such as the rapid formation and dissolution of groups that are in the same geographical area. Ad hoc networks may also refer to geosensor networks, which are distributed ad hoc wireless networks of sensor-enabled miniature computing platforms that are used for the monitoring of phenomena in geographical space, which produce data in real time (Nittel et al. 2004). For example, geosensor networks may monitor water level in a floodplain, or the presence of toxic substance in underground water. By their very nature, which is to support dynamic communication, addressing semantic interoperability issues is indispensable to the functioning of ad hoc networks.

2.4 Background on Semantic Interoperability

Semantic interoperability is a requirement for the meaningful sharing of geospatial data, the integration of different geospatial databases, and the establishment of geocollaboration. According to Bishr (1998), there are several levels of interoperability that can be established between two systems, in order to support the communication between them. Those levels include, among others, network protocol, interoperability at the hardware level, sharing of data files, interoperability between data models, and, at the highest level, semantic interoperability. Semantic interoperability is the knowledge-level interoperability that provides cooperating databases with the ability to resolve semantic heterogeneities arising from differences in the meaning of concepts used to define semantics of data (Park and Ram 2004). According to this definition, databases that can cooperate must be found and the meaning of concepts must be explicit and available.

2.4.1 Problems Related to Semantic Interoperability

In order to achieve semantic interoperability, several issues must be resolved. In this section of the chapter, we review those issues and we emphasize the ones that will be addressed in this thesis.

2.4.1.1 Syntactic, Structural and Semantic Heterogeneity

Semantic interoperability is mainly hampered by semantic heterogeneity (Brodeur et al. 2003). Heterogeneity among geospatial data may be classified as syntactic, structural and semantic (differences in the intended meaning) (Stuckenschmidt 2003; Brodeur et al. 2003). Syntactic heterogeneity occurs when different geospatial database use different formats. The standards that where developed by the Open Geographical Consortium (OGC) aim at resolving syntactic heterogeneity by providing common formats for sharing geospatial data among multiple sources, for example, the Geography Markup Language (GML) which establishes standard geometrical primitives such as GML_Point, GML_Curve, GML_Surface defined in the ISO19107 Standard. In this thesis, we do not address the problem of syntactic heterogeneity. Structural heterogeneity occurs when data is structured differently. For example, when the level of granularity may be different (ex: regions vs countries), or the same real world feature (ex: *lake*) may be represented with different structures, for instance as a class or as the value of an attribute *type of waterbody*. At the spatial level, the same geographic feature may be represented with different geometric primitives (ex: *road* as a line or as a polygon), and at the temporal level, the same event may be represented with different temporal primitives (ex: *inundation* as a date or as a period). Semantic heterogeneity is the difference in the intended meaning of concepts. For example, *geometry of building* may represent the *roof of the building* or the *foundation of the building*. Semantic heterogeneity occurs at different levels, including the metadata level (heterogeneity of metadata describing databases), the database level (heterogeneity of schema of databases) and the data level (heterogeneity of content stored in databases) (Lutz 2005). In this thesis, we do not address the problem of heterogeneity at the data level. Semantic mapping, which is the process of finding semantic relationships or correspondences between database elements (including metadata elements, or database schema elements such as classes, relations and attributes), is a common solution to the semantic heterogeneity problem. However, a major limitation of the semantic mapping approaches is that the quality of their results depends on the quality of their input, or more precisely, on the quality of the representation of semantics (Bakillah et al. 2009). A poor representation of semantics will necessarily lead to poor mapping results, independently of the quality of the semantic mapping process.

2.4.1.2 Semantic Implicitness

Another major but less investigated problem that refrains geospatial databases from achieving semantic interoperability is implicitness. Implicitness means that part or whole of knowledge about a geospatial concept is not made available in a machine readable format and therefore cannot be exploited for semantic interoperability related tasks, including semantic mapping, interpretation of data and assessing whether a data set is suitable for a given need. Implicitness makes differences in the meaning of concepts undetectable (Farrugia 2007), and refrains from identifying similar concepts. Geospatial data is specifically affected by semantic implicitness. Several aspects of geospatial concepts can be implicit. Spatial relations are often implicit (for example, a river that crosses a land parcel) (Lemmens et al. 2006). Context, which is fundamental to compare meaning of different geospatial concepts (Brodeur 2004), can also be implicit in the definition of geospatial concepts. Several definitions of the context have been proposed, and they can be classified into two categories:

- the context as the information that relates an entity of interest to its surrounding environment, for example the geographical location of the entity, the events that have occurred, etc. This definition of context is especially used in context-aware systems (Dey 2001) and applications based on user-profile (Firat et al. 2007)
- the context as a formal representation of a perception of a reality, including view points that express the perspective of an individual or a community of users. This definition of context has been adopted for example in Keßler et al. (2007) who have investigated the impact of context on semantic similarity among geospatial concepts. This definition also includes approaches defining the context as the set of properties of a geospatial concept (Brodeur 2004).

Several researchers in the geospatial database domain consider that multiple spatial representation of the same phenomena can coexist in the same database (Parent et al. 2006;

Bédard and Bernier 2002), meaning that there are different ways to represent the geometry of an object depending on the context. In these approaches, multi-view database models and operators to create view are proposed. But in general, context is seldom explicitly represented in geospatial databases.

In several definitions of concepts, spatial and temporal properties are not explicit but merged into other classes of properties. This means that the separate manipulation of spatial or temporal properties is difficult because they are not explicit and can not be efficiently used for semantic interoperability purposes. Also, most approaches define properties only with their name and range of values, for example "geometry of river" is a "polygon". But this is not sufficient to understand the exact semantics of this spatial property. This is because the polygon may represent "bed of the river'' or ''regions covered with water". The spatial and temporal semantics are therefore implicit. This is a major obstacle to the interoperability of geospatial databases since spatial and temporal properties of concepts cannot be compared.

Finally, another often implicit aspect of geospatial concepts is dependencies between features of concepts. While many concept representation approaches consider concepts as unstructured sets of features (Rodriguez and Egenhofer 2003; Hess et al. 2007; Schwering 2008), most features can be interrelated (ex: the altitude of a city is related to its temperature). Dependencies are useful to understand the nature of real world entities.

2.4.1.3 Real Time

In ad hoc networks, real time is another obstacle to the achievement of semantic interoperability. Since nodes are free to enter or leave the network, a static semantic interoperability approach is not adequate. Real time may have different meanings depending on the context. Real time can refer to the changes that occur in reality and that are immediately stored in the database; in this case, the database is said to be operational (Smirnov et al. 2006). Real time can also be used to indicate real time systems, that is, a reactive system where accuracy of results and short time response are critical (Lambert 2006). According to Kopetz (2011), in a real-time computer system, the correctness of the system behaviour, which is the sequence of output of the system, "depends not only on the logical results of the computations, but also on the physical time when these results are produced." This means that a real-time semantically-interoperable system is a system where the outputs of the components, such as responses to users' queries, are time-dependent. For example, the results of a query that was submitted in a relatively near past could be modified by the arrival of a new agent (with new source of data) in the ad hoc network. In this thesis, we similarly consider that real time refers to the ability of the semantic interoperability solution to adapt to the changes that occur in the network. These changes are: the adding or removal of sources from the network; the formation and dissolution of groups of sources; and other alterations of the groups of sources, including merging two groups, dividing a group into several groups, or removing and adding sources to a group. When such changes occur, the semantic links that were established between the databases of their ontologies could need to be adjusted. From this point of view, real time is an obstacle to semantic interoperability because each change in the network can make obsolete the semantic interoperability strategy that has been deployed. Note however that in this thesis, we do not tackle the issue of reducing the cost of processes to ensure that they can be performed in real time.

2.4.2 Semantic Web and Ontologies

As the volume of data and information available on the Web grow exponentially, the need for solutions to support sharing and reuse of those data and information is growing as well. However, the semantics of available data is often readable for humans but because of their lack of formalism, they are not machine-processable. Therefore, the Semantic Web is presented as a space that supports the exchange of resources and the communication between humans and machines; this would allow the optimal exploitation of large volumes of data and web services. Ontologies are recognized as a major component of a semantic interoperability approach and of the Semantic Web (Agarwal 2005; Curé and Jeansoulin 2009). This section introduces the fundamental concepts related to the Semantic Web and ontologies that are relevant with respect to the research presented in this thesis, which is based on ontologies to describe the semantics of geospatial data.

2.4.2.1 Semantic Web and Standard Languages

The vision of the Semantic Web is represented by Berners-Lee et al. (2001) as an architecture that comprises several layers (Figure 2.1). Syntaxical interoperability is ensured by lower level layers: URI (Uniform Resource Identifier) are used to uniquely identify resources and Unicode is a universal text encoding used to exchange symbols. XML is the eXtensible Markup Language, another fundamental language supporting interoperability at the syntactic level that was proposed by W3C (Bray et al. 2000). It provides a syntax to describe the structure of a document, and its uses namespaces to identify names of tags used in XML documents. XML is not an ontology language; however XML-schemas can be used, to some extent, to specify ontologies. Nevertheless, XML does not support interoperability at the semantic level, since it does not constraint the semantics of the document. The Resource Description Framework Model and Syntax (RFD M&S) supports the description of taxomomies of concepts and properties of concepts. RDF M&S is the first layer that supports interoperability at the semantic level. RDF Schema defines a modeling language on top of RDF. RDF Schema extends RDF by adding more modelling primitives such as domain and range restriction on property (Brickley and Guha 2000). The ontology supports communication between human and machines by introducing semantics in addition to syntax. The ontology also supports the cardinality of relationships between concepts, the transitivity of relationships, inverse relations, etc. Rules support reasoning over data, and can support transformation of data coming from various sources. In our framework, we consider rules as elements that help to refine the definition of concepts in ontologies. In a logic framework, it is possible to infer new knowledge from available knowledge using rules. The Proof layer and the Trust layer enable to verify the validity of statements being made. Signature is used to verify if documents have been altered and Encryption is used to exchange confidential documents.

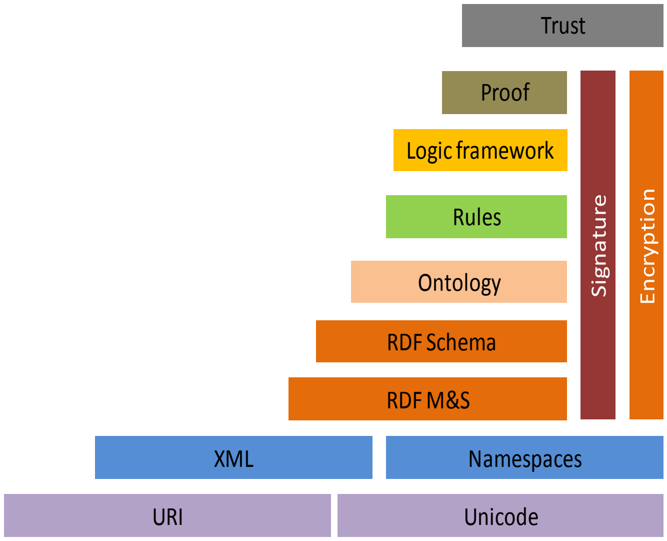


Figure 2.1 Architecture of the Semantic Web (Berners-Lee et al. 2001)

2.4.2.2 Ontologies

In computer science, the most commonly cited definition of ontology is "*an explicit specification of a conceptualisation*" (Gruber 1993). Gruber further explains that a conceptualisation is "*a combination of concepts, and other entities that are assumed to exist in some area of interest and the relationships among them. ... it is a simplified view of the world that we wish to represent for some purpose*." In virtue of this definition, an ontology intentionally omits to represent some aspect of phenomena that are not relevant for the intended purpose. This is why an ontology must be considered as a representation that captures consensual knowledge accepted by a given community (Fensel et al. 2003). In its review on the role of ontologies in GIS, Agarwal stated that the components of an ontology are classes (or concepts), relations, and axioms (Agarwal 2005). Ontologies can be characterized, among other characteristics, by their degree of formality, the extent of explication, the complexity of their structure, and their scope (Lemmens 2006). Those characteristics have an important impact on the quality of semantic interoperability. The degree of formality refers to the formality of language being used to specify the conceptualisation. At the lower level, ontologies that are expressed with natural language are not machine-readable and therefore useless for automated semantic interoperability processes. At the highest level of formality, ontologies are specified in formally defined languages with formal semantics, such as Description Logics (DL). The extent of explication refers to the extent to which the conceptualisation that the user has in mind is made explicit with ontology components. This characteristic is the inverse of implicitness and as such is desirable to improve semantic interoperability. The complexity of the structure is related to the variety of types of ontology components. For example, some ontologies allow only taxonomic (is-a) relations between concepts, while other allow for any kind of relations specified by the ontology designer. The scope of the ontology is related to the area of interest being covered by conceptualisation, also called the universe of discourse. To be interoperable, ontologies must share a non-disjoint universe of discourse.

2.4.2.3 Role of the Ontology in Semantic Interoperability

Ontologies are currently very prominent since their most basic role, which is to represent semantics, is fundamental to every application that must deal with data in a meaningful manner. Basically, ontologies are employed to define the semantics of resources, such a geospatial databases (Brodaric 2007) and the functionalities of geo-services (Lutz 2005; Lemmens 2006). Ontologies support various semantic interoperability tasks, notably source and service discovery. They can be used to provide a description of available sources and services, so that user queries can be matched against those descriptions. Ontologies can also be used to improve query formulation, in order to enhance discovery results. Ontologies are also a major component used to support the integration of data coming from multiple sources (Vaccari et al. 2009). Finally, ontologies are a fundamental building block of the developing Semantic Web (Curé and Jeansoulin 2009). The vision of the Semantic Web is that of an evolved state of the Web where data can be shared, reused and processed across communities of users and producers (Cudré-Mauroux 2006). The Semantic Web was originally defined as "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee et al. 2001). Well-defined meaning is enabled by ontologies that are expressed with formal, standard languages. Research toward developing the Semantic Web have given rise to formal ontology languages and a number of Semantic Web techniques, such as ontology mapping (Kalfoglou and Schorlemmer 2003; Euzenat and Shvaiko 2007).

2.4.2.4 Web-based Ontology Specification Languages

Several ontology languages have been developed, but each of them focuses on different aspects. Some ontology languages have focused on providing an intuitive way of modeling ontologies for humans (such as Ontolingua (Farquhar et al. 1996) and F-logic (Kifer et al. 1995). More recent ontology languages that were developed within the scope of the Semantic Web focus on reasoning capabilities; this is the case of OIL, DAML+OIL and OWL. In this section, we briefly describe the ontology languages that were developed within the scope of the Semantic Web, and finally we focus on Description Logics.

2.4.2.4.1 OIL

OIL (Decker et al. 2000; Fensel et al. 2000) stands for Ontology Inference Layer. OIL was funded by the European Union programme for Information Society Technologies. OIL aimed at providing a general-purpose markup language for the Semantic Web. It is an extension of RDF Schema. It is defined by three layers: Standard OIL, which includes ontology primitives commonly found in ontology languages, Instance OIL, which also comprises individuals (instances) and Heavy OIL, which adds reasoning capabilities. This language provides a predefined set of axioms, such as disjoint classes; however, it does not allow defining arbitrary axioms.

2.4.2.4.2 DAML +OIL

The DAML+OIL ontology language (Horrocks et al. 2002) is the result of the association of the OIL language and the DARPA Agent Modelling Language (DAML). It was also proposed by W3C as semantic markup language for web resources. DAML+OIL is an enrichment of RDF Schema; basically, what DAML+OIL adds to RDF Schema is the possibility to express constraints on the values that a property can have, and constraints of the properties that a class can have.

2.4.2.4.3 OWL

OWL is a semantic markup language that was developed for publishing and sharing ontologies on the web (McGuinness and van Harmelen 2003). In comparison with RDF and RDF Schema, OWL offers improved capabilities because of additional constructs and formal semantics. OWL is based on the Resource Description Framework (RDF) and

Description Logic (DL) Frameworks. OWL allows defining classes with logic connectors such as intersection, union, disjunction, and other restrictions. It defines transitive and symmetric properties, as well as different types of properties (datatype properties and object properties). OWL is composed of three sub-languages with increasing expressivity: OWL-Lite, OWL-DL and OWL-Full. OWL-Lite is the less expressive with only a limited set of constructs available to define classes. OWL-Lite allows defining subsumption hierarchies and simple constraints. OWL-DL adds to OWL-Lite the capacity to express class axioms, Boolean combinations to define more complex classes and arbitrary cardinality. OWL-Full is the most expressive language; it provides meta-modelling capabilities in RDF Schema, however because of its high expressivity, it is undecidable[2]. OWL is the latest recommended language of W3C for publishing and sharing ontologies on the web.

2.4.2.4.4 Description Logics (DL)

Description Logics (DL) are a family of knowledge representation (KR) languages widely adopted for their reasoning capabilities (Baader *et al.* 2003; Lemmens 2006; Fallahi *et al.* 2008). The basic components of DL are primitive concepts, primitive roles and individuals (instances of concepts). Constructors (universal quantification (∀), existential restriction (∃), conjunction, etc.) allow defining complex concepts and complex roles from primitive ones. Common constructors are listed in Table 2.1 (where ∅ represents an empty set, ⊆ represents inclusion, ¬ represents negation, / represents exclusion, ∈ means "element of," the vertical bar | means "where," ∧ means "and, " and → represents an implication).

The semantics of concepts are given by an interpretation $I=(\Delta^I, {}^I)$, where Δ^I is the set of instances and ^{I} is the function that associate instances to their concepts. Therefore, concepts are defined by their sets of instances (or extensions of the concept). Roles are relations between concepts, and their interpretations are sets of relations between instances.

[2] Decidable languages support the expression of problems for which a solution can be computed, i.e., there exists an algorithm that, for a given instance of a problem, is able to compute the correct answer in a finite amount of time.

Table 2.1 Description Logics Syntax

Name	**Syntax**	**Semantic**
Top concept		$\varDelta^I$
Bottom concept		$\varnothing$
Atomic concept	C	$C^I \subseteq \varDelta^I$
Atomic role	R	$\mathrm{R}^{\mathrm{I}} \subseteq \varDelta^I \times \varDelta^I$
Full negation	$\neg C$	$\varDelta^I / C^I$
Concept equality	$C \equiv D$	$C^I = D^I$
Concept inclusion	$C \subseteq D$	$C^I \subseteq D^I$
Concept union	$C \cup D$	$C^I \cup D^I$
Concept intersection	$C \cap D$	$C^I \cap D^I$
Role equality	$R \equiv S$	$R^I = S^I$
Role inclusion	$R \subseteq S$	$R^I \subseteq S^I$
Existential quantification	$\exists R.C$	$\{a \in \Delta^I \mid \exists b.(a,b) \in R^I \wedge y \in C^I\}$
Value restriction	$\forall R.C$	$\{a \in \Delta^I \forall b.(a,b) \in R^I \rightarrow y \in C^I\}$
Maximum number restriction	$\leq NR.C$	$\{a \in \Delta^I \mid \{b \in \Delta^I \mid (a,b) \in R^I \wedge b \in C^I\} \mid \leq n\}$
Minimum number restriction	$\geq NR.C$	$\{a \in \Delta^I \mid\mid \{b \in \Delta^I \mid (a,b) \in R^I \wedge b \in C^I\} \mid \geq n\}$

The subsumption relation (equivalent to the is-a or generalisation/specialisation relationship) writes, for example, as:

$$River \subseteq Watercourse$$

And its interpretation is that all instances of river are instances of watercourse. The following expression gives the example of a concept river whose geometry is a line:

$$River \cap \exists\ HasGeometry.Line$$

It uses the existential quantifier which indicates that every river has at least one geometrical representation, which is a line. The different forms of description logics are determined by the constructors that are used, and define the expressive power of DL. A knowledge base is composed of a TBox and an Abox. The Tbox is the set of terminological axioms; it contains the intentional knowledge. Terminological axioms can be of the form, for example:

$$IntermittentRiver \equiv River \cap \exists\ HasWaterFlow.Intermittent$$

The equality is used to indicate sufficient and necessary conditions to define the concept. The ABox is the set of assertional axioms; it contains the extensional knowledge, that is, the knowledge about individuals. Example of concept assertion and role assertions are as follow:

City (*Montreal*)
IsPartOf (*Montreal*, *Canada*)

An interpretation I (or set of instances) satisfies a TBox T if and only if I satisfies each element of T, that is, all instances belonging to the I set respect the definitions of concepts given in the TBox. When an interpretation satisfies a TBox T, we say that I is a model of T. TBox reasoning processes are the following:

- **Subsumption reasoning**: a concept C is subsumed by a concept D in the TBox T if $C^I \subseteq D^I$ for every model I of T. This means that all instances of C are instances of D.

- **Satisfiability checking**: a concept C is satisfiable in the TBox T if there is a model T where C^I is not empty.

- **Equivalence checking**: the concepts C and D are equivalent in the TBox T if $C^I = D^I$ for every model I of T. This means that all instances of C are instance of C, and conversely.

- **Disjointness checking**: the concepts C and D are disjoint if the intersection of C^I and D^I is empty for all models I of T. This means that C and D have no common instances.

These reasoning processes are useful in establishing semantic mappings across DL ontologies, but they require a priori terminological agreements between ontologies. Abox reasoning allows verifying the consistency of the ABox, meaning that each concept of the TBox can have at least one individual; finding the concept that an individual instantiates; and finding individuals that are instances of a given concept.

2.4.3 Process Calculus: The Lightweight Coordination Calculus (LCC)

Process calculi are another kind of reasoning tool which were developed to describe interactions, message exchange between a set of autonomous agents or processes. In our framework, a process calculus called the Lightweight Coordination Calculus (LCC) is employed to formalize the interactions that occur between nodes of the network during propagation of queries. Process calculus is part of concurrency theory, the theory of parallel and distributed systems in computer science (Baeten 2004). Because they support distributed interactions, they are well adapted to networks of autonomous entities. Process calculi represent interactions with a collection of primitives and operators acting on those primitives. Examples of early process calculus are the Calculus of Communicating Systems (CSS) developed by Robin Milner (Milner 1980), the Communicating Sequential Processes (CSP) (Hoare 1978), and the Algebra of Communicating Processes (ACP) (Baeten and Weijland 1990). The Lightweight Coordination Calculus (LCC) is a more recent process calculus that was developed by Robertson (2004). LCC is a calculus employed to formalize the norms in distributed interactions; it specifies the rules that define message passing between agents. LCC represents interactions as messages between agents. Agents have specific roles that can vary in time; for example, an agent that was a consumer in a given interaction could become a service provider in another interaction. The interactions are governed by social norms; Robertson indicates that the concept of a social norm is employed "to specify behaviours required of agents interacting in a given social context" (Robertson 2004, p. 236). In a generic manner, a social norm is defined by a rule composed of an antecedent and a consequent. The antecedent indicates the agent role and the consequent indicates the behavior of the agent in that role. For example, a social norm can indicate that the role of a data requestor is to send queries. The advantages of LCC in our context include its ability to define interactions without having to determine the features of agents that participate in the interaction, and the fact that it is adapted to a distributed and decentralized environment of autonomous entities such as ad hoc networks. More concretely, in LCC, each agent can make autonomous decisions, based on socially accepted norm that are defined by the *LCC framework*, in parallel with other agents. Agents (denoted A) are defined by their roles and their identifier, while a LCC framework is formed by a set of clauses. Each clause associates with a possible role of an agent the behaviour that the agent can have when he fulfills this role. The possible behaviours include doing nothing, changing roles, sending or receiving a message (denoted *M*). Figure 2.2 shows the LCC syntax.

Framework	*: = {Clause, ... }*
Clause	*: = Agent :: Adef*
Agent	*: = a(Role, Id)*
ADef	*: = null ← C | Agent ← C | Message ← C* *ADef then ADef | ADef or ADef | ADef par ADef*
Message	*: = M ⇒ Agent | M ⇐ Agent*
C	*: = Term | C ∧ C | C ∨ C*
Role	*: = Term*
M	*: = Term*

Figure 2.2 LCC Syntax

Null indicates that there is no message passing. The => symbol indicates message passing between agents, and ← expresses the logic implication. ∧ is the conjunction operator “and”, while ∨ is the disjunction operator “or”. Complex agent definitions can be expressed by using the sequential (then), choice (or), parallel composition (par).

In LCC, an agent *A* that participates in an interaction *I* receives the tuple (*I*, *M*, *R*, *A*, *P*), where *M* is the message, *R* is the role of the agent, *P* is the protocol. The protocol is defined as a collection of LCC clauses, including those that define the protocol framework, the clauses that define the current protocol state, and the clauses specifying the shared knowledge. The protocol framework is static and is unchanged during an interaction. The protocol state is formed by clauses that are constantly modified to keep track of the current protocol state. Shared knowledge is required to carry out a specific interaction protocol. For example, let an agent *A1* be a data producer, which is expressed as *a*(data producer, *A1*), while another agent *A2* is a data consumer, which is expressed as *a*(data consumer, *A2*). Consider that the data consumer’s information need is *X*, which is expressed as InformationNeed(*X*). The action where A2 requests information X from A1 and integrate the received information in his database is expressed with the following LCC fragment:

a(data consumer, *A2*) ::

requests(*X*) ⇒ *a*(data producer, *A2*) ← InformationNeed(*X*) then addToDatabase(X) ← returns(X) ⇐ a(data producer, $A2$).

More recently, LCC was used in a variety of applications, which demonstrate its ability to support dynamic interactions with constraints. Within an approach for semantic integration of geo-services in Spatial Data Infrastructures (SDIs), it was used to specify and reason with the semantics of Web services (Vaccari et al. 2009). The geo-services share explicit knowledge of the interactions in which they participate, and models of interaction are employed to describe the semantics of interactions. LCC was also used for the dynamic verification of trust in distributed and open systems, describing the notion of permission, obligation, and trust (Osman and Robertson 2007). In addition, LCC is employed in multiple stakeholder scenarios, such as in health care, to support medical specialists in sharing clinical knowledge about their patients while respecting policies and rules that pertain to their domain and the patients' confidentiality (Xiao et al. 2009). Those applications have shown that LCC is a useful reasoning language to support interactions while respecting common rules. This is why LCC was chosen to formalize the semantics of interactions that occur between nodes of the ad hoc network when the propagation path for a query needs to be determined according to a selected strategy. The strategy is then formalized as a set of norms that will help a user agent at a node of the network to determine the next query recipient. Depending on specific conditions, a user agent can do nothing (i.e. stop propagation), forward a message (the query), change its role from query recipient to query sender, etc. The query propagation based on LCC is presented in Chapter 7.

2.5 State of the Art on Issues Related to Semantic Interoperability

Semantic interoperability is a complex problem that involves several other specific issues, including that of the discovering of meaningful groups that can collaborate, the issue of knowledge representation and extraction, and the sharing of the meaning across different user communities, in static or dynamic setting. In this section, we present background and existing research related to those issues.

2.5.1 Coalition Discovering

The problem of coalition discovering (which is similar to group discovering, or semantic grouping) can be defined as the problem of gathering members of a network (which can be sources, peers, agents, etc.) into meaningful groups. We consider that there are two main categories of approaches toward group discovering and formation:

- **Agent-based group formation approaches:** those approaches are targeted at multi-agents systems (MAS). They are based on the interactive capabilities of agents. Groups are considered as the result of a negotiation among agents, which is guided by notions taken from game theory, such as gain (what an agent "earns" by entering a group) and payoff (what an agent "pays" or "loose" when entering a group). Examples of such approaches are Dang et al. 2003, Sauro 2005, Oravec et al. 2007, Zheng et al. 2008, Boella et al. 2009, van der Torre and Villata, 2009. Agent-based group formation approaches do not necessarily consider the problem of discovering agents that can be part of the group, but focus on providing the agents with functionalities that enable them to reproduce a negotiation process.

- **Content-based group formation approaches:** in those approaches, groups are formed on the basis of a common knowledge or interest among members. In this thesis, we are mainly interested by those approaches, which are mostly concerned with the problem of discovering nodes of the network that can be part of a group.

The idea of group formation emerged in the context of Peer-to-Peer (P2P) systems with some approaches proposing the concept of interest groups (Giunchiglia and Zaihrayeu 2002), P2P communities (Khambatti et al. 2002; Crespo and Gracia-Molina 2002) and Peer Federations (Bonifacio et al. 2002). Peer groups are useful to structure the search space and therefore support efficient discovery of resources. According to Khambatti et al., P2P communities are sets of peers that share a common interest; this common interest is the intersection of keywords expressing interest of participating peers. P2P communities can also be semantically related peers, that is peers which hold similar knowledge (Crespo and Gracia-Molina 2002; Löser et al. 2003). In Crespo and Gracia-Molina, semantically-related peers are grouped to form Semantic Overlay Networks (SONs), and SONs are organized in a hierarchy (where, for example, a group of peers that are interested in “music” subsumes groups of peers that are interested in “jazz,” “classical,” etc.). There is only one such hierarchy for the whole P2P network. Similarly, in the work of Lumineau and Doucet (2004), a community is defined by a small set of keywords which define an area of interest, and interests are organized in a local hierarchy of interest (formed for a small subset of the P2P network). The difference with the work of Crespo and Gracia-Molina is that there are several hierarchies of interests to describe different local parts of the P2P networks, in opposition to a single global hierarchy common to the whole network. With respect to interest-based groups, SONs gather peers that hold data about similar topics; this is useful to support peer discovery, but it is not a suitable approach for supporting collaboration, since collaboration often implies people holding different but complementary data to work together, not necessarily people just holding similar data. Meanwhile, the existing interest-based peer groups are poorly defined with only keywords.

The concept of collaboration has been recently studied in the area of GIS, emerging into the concept of geocollaboration and collaborative GIS; the latter are designed to support planning and resolution of complex problems involving multiple stakeholders (Balram and Dragicevic 2006; Lee et al. 2006). For example, Balram and Dragicevic have designed a generic collaborative GIS process model that represents the dynamics and patterns of the collaboration process (2006). While it is recognized, in virtue of these approaches, that geocollaboration is a relevant issue; however, collaborative GIS models do not address the problem of forming groups of people that will collaborate. Furthermore, existing group formation approaches are not adequate for the geospatial domain, since the group models are not complex enough to represent spatial and temporal features of groups. Consequently, there is a need to define the notion of groups of geospatial data sources, and an adapted group formation framework.

2.5.2 Knowledge Representation and Extraction

The second issue that is related to semantic interoperability and that will be addressed in this thesis is the problem of knowledge representation and extraction. This is a fundamental issue since all processes that participate in the global semantic interoperability process are based on a knowledge representation. While knowledge representation and extraction is a very large domain in itself, we present the theory and techniques related to our research.

2.5.2.1 Knowledge Representation (KR): Definitions of Concepts

Knowledge representation is the problem of encoding the knowledge that human have about reality, in such a way that it supports reasoning (Kavouras and Kokla 2008). A knowledge representation is not a complete and perfect picture of the reality; but an imperfect abstraction of a portion of reality that is relevant in an application domain. Knowledge representation is a fundamental issue for improving semantic interoperability because it is the support for knowledge sharing (between humans and between machines). Ontologies support knowledge representation. However, ontologies themselves are based on a representation of the concept. In addition, ontology languages allow more or less expressive representation of the concepts. For example, in the lightweight version of Description Logics (DL), concepts can be defined by roles, which represent relations between instances of concepts (for example, the role "HasParent" between instances of the concept "Person"). In more expressive version of DL, additional features such as datatypes (e.g., dateTime) and cardinality constraints (retraining the number of instances that can participate into the relations, e.g. a person can have only two parents) can be expressed. The counterpart of more expressive knowledge representation is the cost and decidability of the reasoning process.

The theoretical basis of knowledge representation approaches depends on the different theories of the concept. From a cognitive point of view, concepts can be defined as mental representation of a category (Medin and Rips 2005) or as explained in Brodeur and Bédard (2001), concepts are the result of the abstraction of a phenomenon in a given context. A category represents a set of real world entities that have similar characteristics, relations, roles, etc. (Kavouras and Kokla 2008). Developing a framework that would guide the assignment of properties to concepts in a universal way is a very difficult task, even if such attempts were made (Margolis and Laurence 1999; Bennett 2005). Usually, the choice of the features defining a concept depends on the purpose or intended task (Brodeur and

Bédard 2001; Tomai and Kavouras 2004). Kavouras and Kokla (2008) define a concept with a term, a set of semantic elements (properties and relations) and their values. This is similar to the concept definition proposed by Schwering and Raubal (2005) where concepts are defined by properties (represented as dimensions in a conceptual space) and property values (represented as values of those dimensions). The properties and relations identified by Kavouras and Kokla include *purpose*; *agent*; *shape*, *size*, *cover*, *property-defined location*; *frequency*, *duration*, *property-defined time*; *is-a*, *part-of* relations; *relative position relations* (upward, downward, behind, etc.); *proximity*, *direction* and *topological relations* (adjacency, connectivity, overlap, etc.); *source-destination relation*. In their approach to measure semantic similarity, Rodriguez and Egenhofer (2003) propose a definition of the concept where features of a concept are classified as attributes, functions (representing what is done to or with an object) and parts (structural component of an object). This classification of properties aims at facilitating the separate manipulation of each type of property. Another set-based concept definition is given by Brodeur and Bédard (2001). They proposed a definition of the concept based on the four-intersection model of Egenhofer (1993). A concept has an interior, defined by its intrinsic properties (e.g. identification, attributes, attribute values, geometries, temporalities, domain), and a boundary, defined by its extrinsic properties (e.g. semantic, spatial, and temporal relationships and behaviours). Some of the previous definitions of the concept assume that the concept is an unstructured set of properties. Bennett (2005) has attempted to give a generic concept definition. He proposed that the properties of an object may be classified as physical (including geometry and material properties); historical (how the object came into existence; the events it has undergone, etc.); functional, including static and dynamic functions; or conventional properties (related to the *fiat* nature of objects). Even if Bennett states that "objects that exhibit one property, will very often also exhibit another property", he does not explicit further those types of dependencies between properties. A second problem is that in most of the definitions, spatial and temporal properties are not explicit but merged into other classes of properties. Consequently, the separate manipulation of spatial or temporal properties is difficult. When knowledge representation is incomplete, a possible solution is knowledge extraction.

2.5.2.2 Knowledge Extraction

As stated in the previous section, semantic implicitness is a problem in geospatial databases as some knowledge about the representation of concepts may be implicit. Knowledge extraction approaches aim at discovering new implicit knowledge from available knowledge, by searching for patterns in data. Knowledge extraction can play an important role in improving semantic interoperability by enriching available knowledge with implicit knowledge. Knowledge extraction includes a range of techniques including *data mining*, *clustering*, *classification*, *semantic information extraction from texts*, *sequential pattern mining*, *association rule mining* and *social network analysis* (Ding and Sundarraj 2007). The idea of using knowledge extraction technique to support semantic interoperability is present in the work of Kavouras and Kokla (2008). They propose a semantic information extraction approach where elements defining a concept are extracted from definitions. First, they perform syntactic analysis (parsing) of definitions; then, they apply rules that locate lexical patterns, to identify some concepts' properties such as location and part-of relations. The extracted elements are used to identify similarities and heterogeneities between geographic categories. In the following sections, we present the knowledge extraction techniques that will be used to generate Multi-View Augmented Concepts (MVAC) (as described in Chapter 5) and discover coalitions (as described in Chapter 4), that is, association rule mining and social network analysis respectively.

2.5.2.3 Association Rule Mining

Association Rule Mining is a kind of Knowledge Extraction for the extraction of association rules from data sets. Association rules are patterns that offer useful information on dependencies that exist between sets of elements (Koh et al. 2007). Association rule mining is widely applied, for example to find correlations in multidimensional data (Ben Messoud et al. 2007) or in XML data (Ding and Sundarraj 2007).

Association rules are logical implication of the form:

$$Head \rightarrow Body,$$

where head is also called the antecedent of the rule, and body the consequent of the rule. The rule indicates that if the head is verified, then the body is also verified. For instance, an association rule (or rule, for short) written in natural language can be:

Road is adjacent to river → flooding risk of the road is high

However, for rules to be processed by reasoning engines, they cannot be expressed in natural language, but they must be expressed to some form of formal syntax and semantics. To keep this discussion general, we do not want to be bound to a language in particular, but express rules with syntax proposed in (Horrocks et al. 2004). In this formalism, body and head are formed with atoms, which can be of the following forms:

- concept axioms: c(x) means that individual x is an instance of concept c
- property atoms: p(x, z) means that value of property p for individual x is z
- relation axioms: r(x, y) means that instance x is related to instance y through relation r

Note that this formalism is the basis of SWRL rules, that is, Semantic Web Rule Language, which is compatible with OWL ontology language.

Rule mining algorithms do not requires particular input form the user, in opposition to other learning algorithms, however they produce large numbers of possible rules that need to be reduce to the set of relevant rules. The selection of relevant, valid rules can be determined with the help of interest measures, which give a quantitative value of the quality of extracted rules (Ceglar and Roddick 2006). Koh et al. (2007) study existing interest measures. For example, the support is an interest measure that tells how many items respects either the head or the body of an association rule, with respect to the total set items. The confidence measures how many items respect the body of the association rule among those that respect the head of the rule. The association rule mining process is generally performed in two steps:

- First, identifying all frequent items in a data set (for example, most frequent values of an attribute). A threshold can be set to determine the most frequent items.

- Second, generate the association rules for the frequent items that meet the requirement on interest measures.

2.5.2.4 Social Network Analysis

Social Network Analysis, or simply put, network analysis, is also a type of knowledge extraction method. The objective of social network analysis is to analyse the "*relationships among social entities, and the patterns and implications of these relationships*" (Wasserman and Faust 1994). Social network analysis includes a set of theories, models, techniques and applications. In recent years, the interest of social network analysis has grown, in social science and in computer science, and especially with the arrival and widely used social networks on the Web. In network analysis, members, or nodes of the network, are viewed as interdependent rather than autonomous, and ties between them are channels for message passing. The network is seen as a source of both opportunities and constraints on its members. The roles of social network analysis that were identified include the following:

- Identify the members or groups of the network that play central role(s);
- Detect information breakdowns, bottlenecks and isolated members or groups;
- Create opportunities to improve information sharing between members and across organizations;
- Improve the efficiency of existing communication links, and highlight the importance of informal communication links;
- Improve the structure and organization of the network;
- Refine communication strategies.

In social science, social network analysis uses questionnaires and surveys to collect information on relationships that characterize individuals from a given group. In computer science, mathematic techniques are employed. Network analysis can be performed from two different points of views: the objective of whole network analysis is to study the structural properties of a network at the global level; egocentric analysis has for objective the study of the network as it is seen from the point of view of one or several of its members (Carrington et al. 2005).

In geographical information science, network analysis has been used in few approaches. Omram et al. (2007) employed social network analysis techniques in support of spatial data sharing (SDS) in spatial data infrastructures (SDI). The authors studied the collective properties of SDS in organizations by mapping the relationships among actors of those organizations using social network analysis. They concluded that for the several organizations that they have studied, spatial data sharing was following the hierarchical relations that were already in place within the organizations.

2.5.3 Semantic Similarity

Semantic similarity is a fundamental notion in GIScience for achieving semantic interoperability among geospatial data, since it allows identifying concepts describing different sources that could answer similar queries. It tells if geospatial concepts are close in meaning, so users of different geospatial data sets can exchange data in a meaningful way. Several semantic similarity models have been proposed in the literature; the models for representing geospatial concepts have been described in a recent review by Schwering (2008). The models are classified as geometric, feature and network models.

Geometric models are based on the concept of multidimensional vector space (or conceptual spaces). Each dimension represents a property (for ex., size); the values of a property (for ex., thin, large) are values of the corresponding dimension. The concepts are represented in this vector space as multidimensional regions. For example, in Figure 2.3, the concept “hill” is defined by a range of values for the properties “height” and “width.”

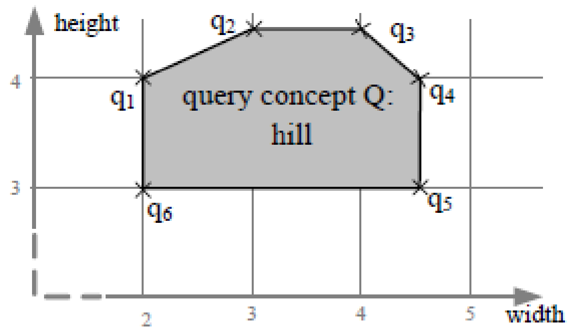


Figure 2.3 Example of concept "hill" represented in a conceptual space (from Schwering and Raubal 2005)

Schwering and Raubal (2005) have proposed a geometric model where the semantic similarity between two concepts is computed as a function of spatial distance between vectors forming the boundaries of the regions representing concepts. The compared concepts must be defined with the same dimensions. In Schwering and Kuhn (2009), this model was extended to take into account relations between concepts. However, dimensions can either match or mismatch, but there is no partial match. In Schwering and Raubal (2005) and Schwering and Kuhn's models, properties are independent of each other; however, Raubal (2004) proposed that dependent properties may be modelled via non-orthogonal dimensions, but this idea was not further formalized.

In network models, concepts are represented as nodes in a graph (Raftopoulou and Petrakis 2005). Figure 2.4 illustrates such a graph of concepts, which are linked with is-a relations.

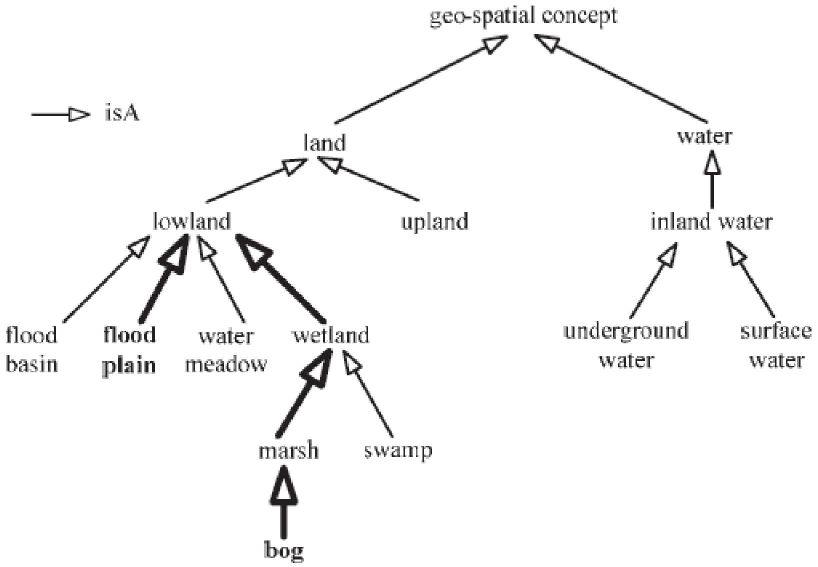


Figure 2.4 Example of graph of concepts used by network models (from Schwering 2008)

Semantic similarity is a decreasing function of the distance (i.e., the number of links) between two concepts. Some network models assign weights to the different types of relationships between concepts (Maguitman et al. 2005). Others combine the shortest path length with the depth of the first common ancestor concept (Li et al. 2003), compare neighbouring nodes of concepts (Do and Rahm 2002), or include the notion of information content (Resnick 1999). The network models often assume a representation of concepts with labels only, while geospatial concepts are more complex.

Feature models, which are based on set theory, represent concepts as unstructured sets of features. The ratio model of Tversky (1977) compares the ratio of common and exclusive features. Rodriguez and Egenhofer's Matching Distance model (2003) combines the ratio model with network distance. The geosemantic proximity model is an example of feature model that determines qualitative relationships among concept (Brodeur and Bédard 2001). It provides geosemantic proximity predicates based on Egenhofer's topological predicates (Egenhofer 1993). Those feature models cannot provide partial matches between features since features either match or mismatch. However, the Matching Distance model has been extended to allow measuring such partial matches (Bakillah et al. 2006). However, those models assume that features of concepts are independent from each other.

The geometric, feature and network models aim at reproducing the human perception of similarity (Schwering and Kuhn 2009). They can be situated at the cognitive level. At another level, we must also mention the logic-based semantic similarity models, which represent concepts with a logical language such as Description Logics (DL). Geometric, feature and network models can be represented with Description Logics (Borgida et al. 2005). For example, d'Amato et al. (2005) proposed a semantic similarity measure for ALC Description Logics. This measure uses instances of concepts. Another logic-based model is Sim-DL by Janowicz (2006). Sim-DL is based on Description Logics (DL) and it compares primitive concepts, roles, and cardinality restrictions on roles, and provides with a weighted sum of similarity with respect to these features. In Sim-DL it is proposed that the weights for the different similarity terms can be computed based on probabilistic methods.

Semantic similarity theories are developed for many purposes, but in many case they are employed for discovering semantic mappings among concepts of different ontologies or database schemas.

2.5.4 Semantic Mapping

Recently, important increases in volume of available data has highlighted the need for achieving semantic interoperability among heterogeneous and multiple sources (Zhao 2007). The manual determination of the semantic correspondences between sources is a time-consuming task; automated approaches are now required in various applications, including geospatial data integration, discovery of sources, query answering, semantic annotation, etc. Semantic mapping is the process of finding semantic correspondences between concepts of different databases, or ontologies of those databases (Euzenat and Shvaiko 2007). The term "ontology matching" is also used to specifically indicate the process of matching ontologies. Two major comprehensive reviews can be consulted (Kalfoglou and Schorlemmer 2003; Euzenat and Shvaiko 2007). Ontology matching takes as input two or more ontologies, (composed of concepts, properties, relations, rules, etc.)

and return the semantic relationships (also called alignment) between ontology components. The semantic relationships are usually based on set theory, and reflect the fact that semantically overlapping concepts necessarily share common instances. The multiple-matching process is the process of matching more than two ontologies. A more formal definition of this matching process is provided by Euzenat and Shvaiko (2007): the multiple matching process is "*a function f, which, from a set of ontologies to match O_1, ...O_n, an input alignment A, a set of parameters p, and a set of oracles and resources r, returns an alignment A' between these ontologies*." The input alignment can be a user-provided or dictionary-provided lexical alignment between terms, for instance. The parameters may be the threshold imposed on semantic similarity to qualify a match. Oracles and resources are external sources of knowledge for the matching, such as lexicon or global, domain or task ontologies. Giunchiglia and Shvaiko (2004) claim that they were among the first to more formally introduce the distinction between syntactic and semantic mappings: syntactic mapping matches names of elements, with linguistic and syntax matching techniques, but without considering semantics. Semantic mapping aim at comparing the meaning of concepts, not only names, and produces semantic relations instead of similarity values. There exists a very high quantity of research on ontology matching. Most of the approaches combine several matching techniques, which were classified by Shvaiko (2004) as follow:

- Linguistic techniques compares terms. Several approaches use linguistic techniques in a preliminary phase (ex: the S-Match algorithm of Giunchiglia et al. 2004, the OWL-Lite Aligner (OLA) of Euzenat and Valtchev, 2004; the FALCON matching system of Hu and Qu in 2008);
- constraints-based techniques, which uses structure of schemas/ontologies to discover matches, including taxonomies, graph of ontology, properties of concepts (Giunchiglia et al. 2004; Hu and Qu 2008)
- Techniques based on auxiliary information, using for example global or domain ontologies and thesaurus (ex: the similarity flooding algorithm of Melnik et al. 2002; the COMA++ system of Massmann et al. 2006).
- Formal matching techniques based on logic reasoning engine, such as S-Match (Giunchiglia et al. 2004) and Ctx-Match (Serafini et al. 2003) based on SAT

(satisfiability) solvers.

In the following, we review the most prominent and representatives approaches to highlight their advantages and limitations.

COMA++ (Massmann et al. 2006) is an extension of the Combination of Matching algorithms (COMA). It's a customizable tool that provides with an extensible library of matching algorithms. Several of those algorithms are based on string matching and languages-based techniques. COMA++ matches with no distinction classes (concepts) and properties. A component is provided to combine the results produced by the different matching algorithms. Each matching algorithm performs differently. COMA++ is able to match SQL, or XML database schemas and OWL ontologies. Its results are quantitative only (similarity value).

Dssim (Nagy et al. 2006) is an ontology mapping system that was designed to produce semantic mapping at runtime, with no human interaction, in a multi-agent framework. Dssim uses the belief function of the Dempster Shafer Theory of evidence to combine different syntactic and semantic similarity results. Dssim is able to match RDF and OWL ontologies, and produces mappings between concepts and between properties. Its results are quantitative only (similarity value).

FALCON (Hu et al. 2007; Hu and Qu in 2008) (**F**inding, **A**ligning and **L**earning Ontologies, and **C**apturing Knowledge by an **ON**tology Driven approach) is also a run-time matching tool. FALCON includes an algorithm that partitions large ontologies into smaller parts and compute mappings between those parts. Then, three elementary matchers are employed to match ontology elements that belong to those parts. The first matcher finds correspondences by using the context of concepts (or other ontology elements) of the ontologies. The context consists in a virtual document, which is a collection of words that are extracted from the concept's description and its neighbouring information. The second matcher is based on a string comparison technique. The third matcher is a structural graph matcher for ontologies. Finally, a central controller manages the matching operations and combines the results of the elementary matchers. The result of FALCON is a similarity value.

GLUE (Doan et al. 2004) is a semi-automatic matching tool whose particularity is to employ learning techniques and a probabilistic approach to determine matches. The GLUE algorithm can be explained as follow. First, the Distribution Estimator component of the GLUE tool takes as input two taxonomies of concepts and their instances. Then, it computes the joint probability distribution between pairs of concepts. The joint probability distribution is related to the number of common instances of the two concepts. Machine learning techniques are used to determine if an instance of a first concept can be an instance of the second concept. Multiple learning techniques can be used, and a meta-learner determines which one should be employed. The types of information used by learners are the names of compared instances, the value formats, and the frequency of values. Finally, the Relaxation Labeller component of GLUE uses heuristics to improve matching accuracy, for example, two concepts are likely to match if their neighbour matches. It can be noted that GLUE depends on the availability and richness of instances.

H-MATCH (Castano et al. 2004; Castano et al. 2003) is an algorithm for dynamically matching ontologies in peer-to-peer systems. H-MATCH uses a peer ontology representation where concepts are describes by properties and relations. The available relations are *same-as*, *kind-of*, *part-of*, *contains*, and *associates*, but H-MATCH cannot compare other types of relations. The algorithm is composed of four matchers that employ syntactic and semantic techniques: the surface, shallow, deep and intensive matchers. The result of H-MATCH is a similarity value and one-to-one, or one-to-many mappings.

The **OWL-Lite Aligner (OLA)** (Euzenat and Valtchev 2004) is also a matching system that aggregates the results of several matchers and whose result is a similarity value. The particularity of OLA is that it was designed to balance the different elements that compose an OWL lite ontology: concepts, properties, names, constraints, taxonomy and instances. The similarity between elements of ontologies depends on the category of ontology elements, and the features that characterise them. OLA employs a variety of distance-based algorithms.

PRIOR is an ontology matching tool founded on **P**rofile p**R**opagation and Inf**O**rmation **R**etrieval methods. The authors of PRIOR argue that the name used to represent the concept is a limited information. Consequently, each concept is enriched with a profile. The notion of concept profile developed by PRIOR is similar as that of the FALCON virtual document. The profile of a concept is given by its name, labels, comments, restrictions on values of properties and other descriptive information. Linguistic and structural information is used to map profiles, with cosine similarity.

RiMOM (Tang et al. 2006; Zhang et al. 2006) is an ontology matching tool that aims at determining optimal mappings by combining different matching strategies that exploit different information. First, given two ontologies, the structure and label similarities are computed. Then, if the label similarity value is the highest, RiMOM system employs linguistic matching strategies. Otherwise, if the structural similarity is the highest value, strategies based on similarity propagation with respect to structure are executed. The results of different strategies that are executed independently are combined with a linear interpolation method.

Similarity Flooding (SF) (Melnik et al. 2002) is a matching algorithm based on the notion of similarity propagation. The principle of similarity propagation is that two ontology elements (e.g. concepts) are similar if their neighbors are similar, so the similarity "propagates" to adjacent elements. In the SF matching algorithm, database schemas are represented as directed and labeled graphs, and syntactic matching techniques are employed to compute a preliminary mapping.

Quick Ontology Mapping (QOM) (Ehrig and Staab 2004) is a mapping approach that focuses on reducing the complexity and cost of the matching process. It aims at achieving a trade-off between quality of results and efficiency of the computation process. QOM takes as input RDF ontologies with concepts, properties, relations, axioms, and instances. QOM does not compare all pairs of nodes between the RDF trees to avoid costly processes. It first computes preliminary mappings based on lexical knowledge, and then iterates to determine mappings based on the structure of ontologies. The result of QOM is a similarity value.

Ctx-Match (Serafini et al. 2003) is dedicated to the matching of OWL-DL ontologies. A context is a model that is valid for a given community. Ctx-Match is able to retrieve semantic relations between concepts: equivalence, subsumption, intersection and disjointness. It is based on the theory of contextual reasoning in AI. Ctx-Match performs a contextualization phase, where the focus of a concept, which is the set of ancestors of the concept in the hierarchy, is determined. Ctx-Match then encodes the problem of finding relations between concepts of different ontologies in a problem of logic satisfiability. WordNet is also used as an external resource to enrich the mapping process. Together with S-Match, Ctx-Match is one of the very few approaches that produce qualitative relationships.

S-Match (Giunchiglia et al. 2004) was the first system to implement a semantic mapping approach, and not only a syntactic mapping approach. It takes two graph-like structures as input, and for each node determines the concept as the node, which contains a conjunction of all possible senses of the term at the node, derived from WordNet, plus the ancestor nodes. The concept at the node is then expressed as a propositional logic formula. For each pair of nodes of different ontologies, each type of semantic relationship (equivalence, more general, less general, overlap and mismatch) is verified with a satisfiability (SAT) solver. S-Match uses several string-based matchers and structural matchers. S-Match does not consider attributes of concepts and therefore is not suitable in the presented form for geospatial databases.

Table 2.2 summarizes the characteristics of the above-mentioned matching approaches. This table aims at comparing approaches with respect to (1) the type of input taken by the matching process, (2) the mapping processes and techniques employed, at the element level, and at the structure level, and (3) the type of output. The type of input indicates the expressivity of knowledge representation that approaches can take into account. Most approaches are intended to compare database schemas or ontologies that are expressed in a given format or standard, in order to be automatic or semi-automatic. Most, but not all approaches are able to cope with concepts having attributes and properties. However, several, such as H-Match and S-Match, consider only some types of relations. Very few, such as OLA, take into account constraints. Most important to the geospatial domain, none considers spatial and temporal properties and relations in an explicit manner. It can be argued that spatial and temporal properties can be considered as any other properties, but in this research we want to outline that spatial and temporal properties must be represented with more complex structures than thematic properties.

Concerning the mapping processes and techniques employed, most of the approaches employ more or less sophisticated syntactic and linguistic techniques at the element level. However, at the structure level, besides using taxonomic relations (ex: S-Match, QOM, GLUE, FALCON, and many others) and neighbor information (ex: OLA, RiMOM), it is visible that less matching criteria have been developed. This means that existing matching approaches are less suitable for ontologies with more complex structures, such as for geospatial databases.

At last, with respect to the type of output, only Ctx-Match and S-Match issue qualitative relations, while the result of other approaches is a similarity value. The advantage of qualitative relations with respect to quantitative similarity is that the former are more expressive and allow more useful interpretation of semantic mappings. For instance, a similarity of 0,50 between two concepts is less useful than "concept A includes concept B", because the inclusion allows to determine that all instances of B are instances of A as well. This is very useful indeed to determine if queries posed to both concepts are comparable, which is fundamental for semantic interoperability. For comparison purposes, the last column of Table 2.2 contains the characteristics of the G-MAP semantic mapping model that we have developed and presented in this thesis.

Table 2.2 Comparison of Matching Approaches

			SF	**GLUE**	**OLA**	**QOM**
Type of input	**Format of ontology**		Relational and XML schema	Taxonomy, Relational and XML schema	RDF and OWL ontologies	RDF ontologies
	Used Ontology Elements		Concept, relations and attributes	Instances, Concept, Attributes, Values	Concept, Properties, Is-a relations, instances, constraints	Concept, Properties, relation, taxonomy, instances, axioms
Mapping processes and techniques	**Element Level**	**Syntactic**	String matching, Datatypes, keys	Domain constraints	String matching, language techniques, datatypes	String matching, language techniques
		External resources	WordNet	Mapping reuse	WordNet	-

	Structure Level	**Syntactic**	-	Taxonomy	Taxonomy, Neighbours, Iterative fixed point computation	Taxonomy, axioms
		Semantic	Propositional SAT reasoner	-	-	-
Type of output	**Mapped Elements**		Concepts, attributes	Concepts	concepts	Concepts, relations, instances
	Type of relations		Similarity value	Similarity value	Similarity value	Similarity value

Table 2.2 (continued)

			S-MATCH	**FALCON**	**COMA++**	**CTX-MATCH**	**DSSIM**
Type of input	**Format of ontology**		XML schema, OWL ontologies, Taxonomy	RDF and OWL ontologies	Relational Schema, XML schema, OWL ontologies	Taxonomy, OWL ontologies	RDF and OWL ontologies
	Used Ontology Elements		Concept, taxonomy	Concept, properties, relations, taxonomy	Classes, properties, attributes, datatypes, aggregation and specialization relationships, instances	Concepts, taxonomy, attributes	Concepts and properties
Mapping processes and techniques	**Element Level**	**Syntactic**	String matching, language techniques	String matching, language techniques	String matching, language techniques, datatypes	String matching, language techniques	String matching
		External resources	WordNet	WordNet	Thesaurus, Mapping reuse	WordNet	WordNet
	Structure Level	**Syntactic**	-	Graph , Structural affinity	Taxonomy	Taxonomy	graph
		Semantic	Propositional SAT solver	-	-	Propositional SAT solver	-
Type of output	**Mapped Elements**		concepts	Ontology entity	Concepts and properties	Concepts	Concepts and properties
	Type of relations		equivalence, more general, less	Similarity value	Similarity value	equivalence subsumption, disjointness,	Similarity value

			general, mismatch, and overlapping			and intersection	

Table 2.2 (continued)

			H-MATCH	**PRIOR**	**RIMOM**	**G-MAP (our approach)**
Type of input	**Format of ontology**		OWL ontologies	RDF and OWL ontologies	OWL ontologies	XML schema, OWL ontologies
	Used Ontology Elements		Concepts, properties, relations, values	Concept profile : Name, label, comment, property, restriction, description	Names of concepts, properties, taxonomy, instances	Concepts, properties and their values, relations, mixed properties and relations (ex : spatio-temporal, spatio-thematic), dependencies, spatiotemporal descriptors, context of concept
Mapping processes and techniques	**Element Level**	**Syntactic**	Language techniques	String matching, Language techniques	String matching	String matching, Language techniques
		External resources	Common thesaurus	-	WordNet	WordNet, OpenCyc spatial properties and relations, OpenCyc temporal ontology
	Structure Level	**Syntactic**	-	Graph, taxonomy	Taxonomy, Similarity propagation	Taxonomy, dependencies and other concepts features
		Semantic	-	-		Rule-based reasoning engine

Type of output	**Mapped Elements**		concepts	concepts	concepts	Concepts, views, and other concepts features
	Type of relations		Similarity value	Similarity value	Similarity value	equivalence, includes, included in, strong overlap, weak overlap, disjoint

Based on the above comparison of approaches, we identify some requirements that a semantic mapping approach for ad hoc networks of geospatial databases should meet:

- the approach should generate semantic mappings automatically;
- the approach should consider spatial, temporal and spatiotemporal properties in a separate and explicit manner;
- the approach should take into account the different contexts of a concept, and the variation of definition of the concept under those different contexts;
- the approach should provide additional reasoning criteria and rules to use the complex structure of geospatial concepts, including dependencies between features defining the concept.

In addition to those characteristics, the constraints of ad hoc networks require that the semantic mapping process should be triggered at the right time, i.e., following the issuance of a query, to match the right ontologies. This is the object of query propagation techniques.

2.5.5 Query Propagation

One of the advantages of networks is the availability of a very large number of sources. However, it is also expected that one can retrieve the data he or she needs without having to access a large number of sources, which would be time consuming and ineffective. The goal of query propagation is to find the sources to which a given query should be forwarded. Because networks are significantly large, existing query propagation approaches usually assume that the network is decentralized, meaning that there is no central "authority" that has a global knowledge of the network; rather, the query is transitively propagated from one source to another in a decentralized manner (Cudré-Mauroux 2006; Montanelli and Castano 2008). Each intermediate recipient of the query is responsible for determining to which of its neighbour it will forward the query. In this thesis, we define the query propagation problem as follow:

Query propagation problem: for a query issued by a node of the network, determine to which nodes this query should be propagated, and in what order, in order to obtain the optimal query results. The result of the query propagation algorithm is an oriented propagation graph, where nodes are the selected sources and directed arcs between nodes are paths along which the query is routed (Figure 2.5). The nodes that are closest to the requestor node in the graph are the more relevant to answer the query. As we move further along a path, the relevance of nodes decreases.

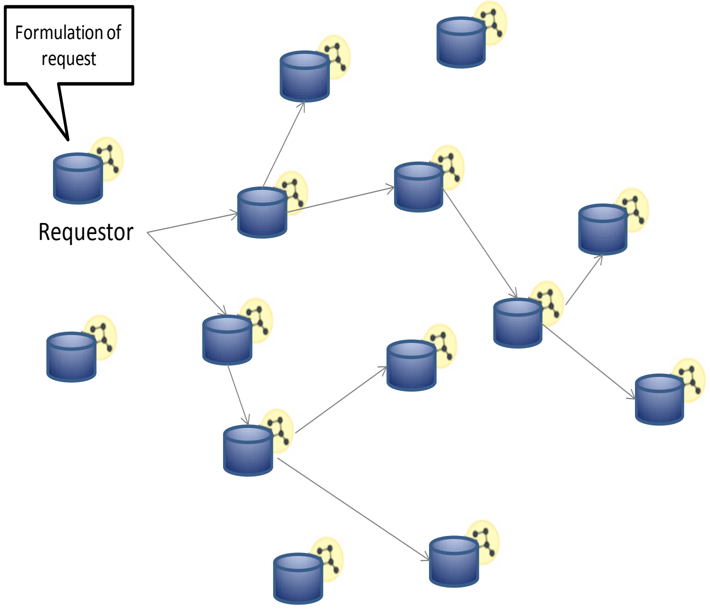


Figure 2.5. The problem of determining the query propagation graph

In this thesis, we add an additional constraint to the problem of query propagation. Many approaches for query propagation rely on semantic mappings between the query and ontologies of databases at nodes to determine the propagation path (see Table 2.3). We consider that the determination of the query propagation graph should not rely on semantic mappings between concepts of databases' ontologies. Computing semantic mappings between the query and every concept of every ontology is very costly. Rather, we propose to select the relevant nodes based on a more global criterion. Then, when the relevant nodes are selected and the propagation graph is computed, semantic mappings between the query and the ontologies of selected nodes can be computed in order to retrieve the more relevant concepts. This will ensure that we compute only the minimal number of semantic mappings. This aspect will differentiate our approach from existing ones. Our approach for query propagation is presented in Chapter 6. In the following, we review and compare the representative approaches to highlight their strengths and limitations. In Table 2.3, the chosen approaches are described according to six characteristics that we have identified. We have identified these characteristics by analysing the differences between the various query propagation approaches:

- the type of networks for which the approach was mainly dedicated;
- the format or description used to represent knowledge on nodes of the network;
- the formalization of queries; the more the query is rich and complete, the more the query propagation approach may be accurate. This, however, also have an impact on the cost of the propagation;
- the criteria for selecting query recipients (nature of the information used to propagate queries);
- whether it is assumed that semantic mappings are already computed, or if they are computed at run-time;
- whether there is an update mechanism that reacts when the network is modified (for example, when a node is added to the network).

Most query propagation approaches were designed for peer-to-peer networks. In several approaches, peers can hold an ontology or a formal description expressed in a language such as XML or OWL. However, some approaches did not commit to a specific language and the knowledge held by peers is simply represented by a set of keywords indicating the expertise of peers, for instance (Haase et al. 2008). In some of the approaches listed in Table 2.3, the knowledge held by peers is related to their area of expertise (e.g., an ontology describing the data stored in a peer's database). In the Intelligent Search Mechanism (ISM) proposed by Zeinalipour-Yazti et al. (2005), each peer holds a list of the previous queries that this peer was able to answer.

Similarly, queries can be represented with a formal language or as a simple keyword or set of keywords. In the REMINDIN' (Routing Enabled by Memorizing Information about Distributed Information) approach proposed by Staab et al. (2004) and in Edutella (2004), queries are expressed as RDF statements. In the approach of Giunchiglia and Zaihrayeu (2002), SQL queries are used. The advantage of RDF or SQL queries over keyword is the increased expressivity that will improve the ability of the approach to retrieve relevant peers.

In the majority of the approaches, the criterion for selecting recipient peers is a kind of correspondence between the query and the knowledge held by peers. For instance, in the approach of Haase et al. (2008), a shared ontology is employed to describe the expertise of each peer using a common vocabulary. The selected recipients of a query are determined based on the semantic similarity between the query subject and the expertise of peers. Other approaches use semantic mappings between query concept and concepts of peers (for example, Semantic Link Networks (P2PSLN) approach developed by Zhuge et al. in 2004). In the H-Link query routing algorithm, the H-Match semantic similarity model (Castano et al. 2003, 2004) is used to find mappings between ontologies of peers and to form a semantic overlay that supports query propagation (Montanelli and Castano 2008). In other approaches, groups gathering peers having common characteristics are employed. In Giunchiglia and Zaihrayeu (2002), the notion of interest group was introduced with the purpose of determining, for a given query, the query scope, that is, the set of nodes a query will be propagated to. This means that a query can be propagated only inside one group. In the Edutella approach (Nejdl et al. 2004), super-peers are responsible for distributing queries to appropriate subsets of peers. When a peer enters the network, it registers to a super-peer and provides its metadata, which describes the most significant features of this peer. While the advantage of using groups is to improve the scalability of the approach (Nejdl et al. 2004), propagation approaches based on groups tend to restrain the propagation to a predetermined set of peers, which may reduce accuracy.

Finally, another category of approach selects the query recipient based on the queries that a peer has previously answered. The REMINDIN' approach allows a peer to evaluate the confidence value of another peer based on the number of correct answers given by this peer for a query. The confidence value is updated when the peer has new interactions with respect to the same topic of interest. The query is propagated to the peers who have the highest confidence value. In the Intelligent Search Mechanism (ISM) proposed by Zeinalipour-Yazti et al. (2005), when a peer receives a query, it retrieves the answers to previous queries in order to select only the peers that are the most likely to give a relevant answer to this new query. A query similarity function is employed to compare past queries to the current query. The query is then forwarded to selected peers. Note that the performance of those interaction-based approaches depends on previous queries, so they may not perform well at the beginning, or have poor performance when a rarer query is submitted.

Few of the approaches include an update mechanism that takes into account the changes that may affect the network. When a peer enters or leaves the network, the semantic links are updated in P2PSLN, and the knowledge being held by super-peers in Edutella (which consist in routing indices) is updated too. However, none of the approaches evaluate whether the answer to a recently answered query can be reconsidered when a new source enters the network.

Table 2.3 Query Propagation Comparative Study

	Cudré-Mauroux 2006	**REMINDIN'**	**P2PSLN**	**ISM 2005**	**Mandreoli et al.**	**H-Link**
Type of network	Peer-to-peer	Peer-to-peer	Peer-to-peer	Peer-to-peer	Peer-to-peer	Peer-to-peer
Knowledge representation	No specific language, concept and properties (RDF, XML)	RDF statements	XML schema	list of the most recent past queries	Concepts as in an ontology, a relational table or XML schema	OWL ontologies with concepts, properties, datatype properties
queries	Concept and its properties	SeRQL query language	Concept and its properties	Set of keywords	A concept	Concept and properties
Criteria for	syntactic	Confidence	Semantic	Similarity	Semantic	Semantic

selecting query recipients	and semantic criteria measuring the quality of mappings	between peers based on previous correct answers given by a peer for a query	mappings	between current and past queries	Mappings	affinity based on semantic mappings
Semantic mappings	Already computed	Already computed	Computed at run-time	Computed at run-time	Already computed	Computed at run-time
Presence of update mechanism	no	no	Update of semantic links if arrival or departure of peer	no	no	no

Table 2.3 (continued)

	Giunchiglia and Zaihrayeu 2003	**Haase et al. 2008**	**GNutella 2004**
Type of network	Peer-to-Peer	Peer-to-Peer	Peer-to-Peer
Knowledge representation	Concepts and attributes of a database schema	Set of keywords describing source expertise	RDF schemas
Representation of queries	SQL queries	Set of keywords	RDF statements
Criteria for selecting query recipients	Interest groups	Semantic similarity between query subject and peers expertise	Super-peers
Semantic mappings	Already computed (semantic correspondences)	Computed at run-time	Already computed
Presence of update mechanism	no	no	Update indices stored by super-peers

Because the approaches were mainly targeted at thematic-only peer-to-peer networks, they do not take adequately into account the geospatial aspects of the queries and sources. In addition, in each of them, a unique criterion for selecting query recipients is employed, while different criteria might be useful in different situations.

2.6 Discussion

Semantic interoperability is the process that supports the meaningful sharing of data between users of multiple geospatial databases. Semantic interoperability can be achieved when heterogeneity and implicitness are overcomed. Ontology is therefore a major component of a semantic interoperability framework since it represents semantics; semantics mappings act as bridges between different abstract representations of a common reality, and consequently, are a viable solution to resolve semantic heterogeneities. However, new challenges arise when we consider semantic interoperability in ad hoc networks, and traditional solutions need to be reconsidered. The literature review demonstrates that new approaches for semantic interoperability in dynamic networks have emerged in the computer science domain; however, these approaches are mainly concerned with performance and cost issues, and therefore, most use a simplified representation of semantics. In addition, the results provided are mostly similarity values, which are less expressive than qualitative, semantic relationships. We argue that while performance and cost are real issues, a poor representation of semantics is not suitable for the geospatial domain, where semantics of spatial and temporal features need to be represented as well. As a counterpart, the cost of computing semantic mappings can be reduced by selecting only the minimal number of sources that must be reconciled to answer a query. This is the role of the query propagation approach in our framework. This literature review shows that improving knowledge representation and providing an appropriate semantic mapping approach is a requirement to meet the constraints of the geospatial domain and those of ad hoc networks.

CHAPTER 3

A Conceptual Framework for Real Time Semantic Interoperability in Ad Hoc Networks of Geospatial Databases

M. BAKILLAH, M.A. MOSTAFAVI

Submitted to *International Journal of Geographical Information Science*

3.1 Résumé de l'article

Les récentes avancées technologiques ont permis le développement des réseaux ad hoc, un ensemble ouvert de bases de données géospatiales qui peuvent s'auto-organiser en temps réel pour répondre à des besoins ponctuels. Bien que plusieurs approches pour l'interopérabilité sémantique aient été proposées, un cadre conceptuel qui s'intéresserait aux problèmes soulevés par ce type de réseau n'a pas été proposé. Le manque de sémantique explicite, en particulier, la sémantique de la spatialité et de la temporalité, ainsi que l'absence d'une approche de réconciliation sémantique qui soit en mesure de prendre en compte une riche représentation de la sémantique et qui permet l'interaction et la collaboration dynamique entre les bases données géospatiales pertinentes du réseau comptent parmi les limitations majeures des approches existantes. Afin de répondre à ces problèmes, cet article propose un cadre conceptuel pour l'interopérabilité sémantique en temps réel dans un réseau ad hoc de bases de données géospatiales, lequel se base sur les principes des réseaux sociaux. Le cadre conceptuel comprend trois principaux composants. Le premier de ces composants est un modèle conceptuel représentant les coalitions dynamiques de bases de données géospatiales, qui ont pour but de permettre la géo-collaboration. Le second composant est le modèle du Concept multi-vues augmenté (MVAC), lequel a été développé pour fournir une riche représentation de la sémantique. Le troisième composant est un nouveau modèle de mapping sémantique, G-MAP, un modèle automatique qui compare les concepts MVAC par un processus automatique qui utilise des

moteurs d'inférence basé sur de nouveaux critères de matching. Le cadre conceptuel a été implanté afin de démontrer qu'il permet d'améliorer l'interopérabilité sémantique entre les bases de données géospatiales.

3.2 Abstract

Recent technological advances have enabled the development of ad hoc networks, where an open collection of geospatial databases may self-organize in real time for short-term needs. While several semantic interoperability approaches have been proposed, a framework that address issues related to such networks is missing. The lack of explicit semantics, especially semantics of spatiotemporal features, and the lack of a semantic reconciliation approach that takes into account a rich representation of semantics and that supports dynamic interaction and collaboration between appropriate geospatial databases of the network are some of the major limitations. To address these issues, we propose a real time semantic interoperability framework for ad hoc networks of geospatial databases founded on social network principles. The framework includes three main components. The first component is a conceptual model for dynamic coalitions of geospatial databases designed to support geocollaboration. The second component is the Multi-View Augmented Concept (MVAC) model, designed to provide a rich representation of semantics. The third component is a new semantic mapping approach, the G-MAP, which compares MVAC concepts in an automatic manner, using inference engines based on new matching criteria. The framework was implemented to demonstrate that it improves semantic interoperability among geospatial databases.

3.3 Introduction

Developments of networking and communication technologies have allowed a shift from isolated geospatial databases to ad hoc networks, where databases can freely join or leave the network, and form groups to share data and services. For the geographic information (GI) community, these developments resulted in an increased availability of geospatial data. This was supposed to support the reuse of data distributed over different geographical information systems (Lutz et al. 2003; Lemmens 2006). However, these developments are not sufficient since meaningful sharing of data requires semantic interoperability (Lutz et al. 2003; Kavouras et al. 2005; Bian and Hu 2007). The issues posed by this context are various and complex. On the one hand, it requires resolving heterogeneities caused by different and implicit representations of the spatial, temporal and thematic aspects of concepts (Schwering 2008). Many semantic interoperability approaches have focused on the thematic aspects, leaving aside the explicit representation of spatial and temporal aspects. On the other hand, in ad hoc networks, the geospatial databases that have to interoperate are not known in advance (Lutz et al. 2003). Several semantic interoperability approaches for networks have focused on automating the semantic mapping process, but this is not the only concern. The structure of an ad hoc network is continuously modified by network events such as the addition or the removal of a source, formation of groups of geospatial databases, etc. It is essential to consider those issues for real time semantic interoperability of geospatial databases in an ad hoc network (Keeney et al. 2006). This paper proposes a conceptual framework for real time semantic interoperability among geospatial databases of ad hoc networks that will address such issues. We argue that real time semantic interoperability can be represented as communication in social networks. Nowadays, many organizations (e.g. government, enterprises) depend increasingly on others because they need to increase their knowledge to reduce risk of wrong decisions (Fox 2008). In the GIScience community, several projects are conducted with the cooperation of different organizations (Balram and Dragicevic 2006; Staub *et al.* 2008). For instance, disaster management crosses human-defined boundaries. In addition, emerging environmental models designed to solve complex problems incorporate multiple models developed by collaborating groups (Bian and Hu 2007). The traditional paradigm where organizations were centralized and all the data was stored in a single geospatial database is no longer appropriate (Aberer et al. 2004). The emergent vision in GIScience is that of self-organized, autonomous, and networked organizations, each offering and requiring geospatial data and services at different times, a vision close to that of the Geospatial Semantic Web. Such networks can be qualified as social networks. In this paper, we identify properties of social networks and use them to set the guidelines of our framework. The conceptual framework is formalized as a conceptual model representing the fundamental elements of real time semantic interoperability for geospatial databases and their interactions. It includes a novel model for coalitions of geospatial databases as a mean to support collaborative semantic interoperability and gather users of geospatial databases that can work toward a common goal. A central component of the framework is a novel geospatial concept representation, the Multi-View Augmented Concept (MVAC) model. The role of this model is to provide a richer representation of semantics that will improve semantic interoperability by making explicit some aspects of concepts that are often implicit, including contexts, semantics of spatiotemporal properties of concepts, and dependencies among the concept's features. This conceptual representation is used by the G-MAP automatic semantic mapping approach, which, in opposition to most existing methods, explicitly considers the thematic, spatial and temporal perspectives that characterize geospatial concepts. We also outline different real time strategies to propagate geospatial queries to relevant sources of the network. The framework was tested in the last section of this chapter.

3.4 Moving from “Semantic Interoperability” to “Real Time Semantic Interoperability in Ad Hoc Networks of Geospatial Databases”

Semantic interoperability ensures meaningful data sharing among different geospatial databases (Agarwal 2005; Bian and Hu 2007). It is the knowledge-level interoperability that provides cooperating databases with the ability to resolve semantic heterogeneities arising from differences in the meaning of concepts (Park and Ram 2004). While there are many approaches that aim at enabling semantic interoperability, new problems arise when we consider ad hoc networks of geospatial databases, in opposition to “traditional” semantic interoperability between a static, small set of known databases. Geospatial databases that are developed by different communities and for different purposes are affected by semantic heterogeneities that prevent them from interoperating (Brodeur et al. 2003). To make meaningful data exchanges, we must solve heterogeneities caused by thematic differences but also by different representations of the spatial and temporal properties of concepts (Brodeur et al. 2003; Schwering 2008). Resolving semantic heterogeneities requires addressing two complementary issues: (1) representing semantics of geospatial data, and (2) discovering semantic relationships among geospatial concepts describing geospatial data.

Semantics is the meaning of expressions in a language (Kuhn 2005). Ontologies, which are explicit specifications of a conceptualization (Gruber 1993), are often used to represent the semantics of data (Brodeur et al. 2003; Kuhn 2003; Rodriguez and Egenhofer 2003; Agarwal 2005; Fonseca et al. 2005; Kavouras et al. 2005; Arpinar et al. 2006; Klien et al. 2006; Lemmens 2006; Brodaric 2007). Concretely, ontologies are composed of concepts, relations, properties and axioms that represent a domain of interest (Agarwal 2005). Fonseca et al. (2002) proposed an ontology-based framework, called ontology-driven geographic information system (ODGIS), to resolve heterogeneity of geographic data. Each ontology is a component that describes the view of a given geospatial information community (GIC). By browsing through ontologies, the users access information available in the embedded knowledge of the system. This approach is suitable for a limited number of sources only, because there is no discovery system in the proposed architecture. In other approaches for semantic interoperability among geo-services, ontologies are used to describe the functionalities of geo-services (Klien et al. 2006; Fallahi et al. 2008; Vaccari et al. 2009). However, ontologies alone are not the complete solution, since they are themselves semantically heterogeneous (Vaccari et al. 2009) and their degree of semantic explicitness is varying (Obrst 2003). In the ontology development process, some knowledge may have been left implicit (where implicit, in this context, also means that the knowledge is not machine-understandable). For instance, geospatial concepts and their spatial and temporal properties are often described by definitions (for example, “floodplain is a meadow that is adjacent to a river”). Definitions cannot be exploited directly to compare concepts, since they are expressed in natural language (i.e., not expressed with a machine-readable language) (Kavouras and Kokla 2008). Other implicit knowledge about geospatial concepts include the context of the concept (for example, *river* in the context of *flooding* or *dryness*, *navigation* or *swimming*, etc.), and dependencies between features of geospatial concepts (for example, the geometrical representation of a river is related to its width). Leaving knowledge implicit makes differences in meaning undetectable (Farrugia 2007). To detect such differences, a representation of semantics which incorporate all needed elements is required.

Discovering semantic relationships among geospatial concepts describing different databases is the goal of *semantic mapping*. Many semantic mapping frameworks aim at discovering quantitative relationships (a semantic similarity value), and some aim at discovering qualitative relationships (usually based on set theory: equivalence, inclusion, overlap, etc.). Some semantic similarity frameworks are dedicated to the comparison of geospatial concepts (Rodriguez and Egenhofer 2003; Schwering and Raubal 2005; and see the review of Schwering 2008). Nevertheless, in general, these models do not consider semantics of spatial and temporal properties of concepts in an explicit manner. Therefore, they cannot detect spatial or temporal similarities explicitly. The G-Match model of Hess et al. (2007) considers that two geospatial concepts with different geometric primitives (for ex: point vs. line) are geometrically different and two geospatial concepts having the same geometric primitive are geometrically equivalent. However, the same geometrical primitive (e.g. surface) may represent two different parts of the same object that are semantically different. For example, two concepts "*house*" may refer to the same real world phenomena, but their spatial extent refer to different things, e.g. the house's roof or house's foundation. In the broader domain of semantic matching between ontologies or database schemas, existing semantic matching systems often integrate several matching techniques, which were classified by Shvaiko (2004). The proposed classification identifies linguistic techniques, which compare terms. Several models integrate linguistic techniques in the pre-processing phase: for example, S-Match (Giunchiglia et al. 2004), OWL-Lite Aligner (OLA) (Euzenat and Valtchev 2004), and FALCON (Jian et al. 2005). Constraint-based techniques use structure of schemas/ontologies to discover matches, including taxonomies, graph of ontology, properties of concepts (Giunchiglia et al. 2004; Jian et al. 2005). Techniques based on auxiliary information use external resources, for example, global ontologies or thesaurus (Melnik et al. 2002; Massmann et al. 2006). Formal matching techniques are based on logic reasoning engines, such as S-Match (Giunchiglia et al. 2004) and Ctx-Match (Serafini et al. 2003), which are based on SAT (satisfiability) solvers. Among these approaches, there are none which use the dependencies between features of concept as valuable structures to discover semantic mappings. This may be because dependencies are seldom represented in concept definitions. In our framework, we will investigate how these structures can improve the semantic mapping process.

However, the resolution of semantic heterogeneities is not the only issue that must be addressed to ensure real time semantic interoperability in dynamic and open environments, such as ad hoc networks. Because of the significant number of sources, we cannot assume that geospatial data consumers know which sources are relevant to their current needs. Therefore, a solution to structure the network (at the semantic level) into groups of databases is needed. Such a semantic structure would support discovery of relevant sources, and propagation of users' queries to these sources. While there exist semantic grouping approaches for networks (Crespo and Garcia-Molina 2002; Khambatti et al. 2002; Montanelli and Castano 2006), these approaches were developed for generic peer-to-peer systems and are not adapted to ad hoc networks of geospatial databases. Furthermore, our goal is to demonstrate how the creation and management of groups of geospatial databases is integrated into the global framework for real time semantic interoperability in ad hoc networks of geospatial databases. In addition, to address the real time aspect of the semantic interoperability process, we need to consider that the network is open and that the groups that are formed in the network are dynamic; as a result, the semantic interoperability process must constantly adapt to the changes. While some semantic interoperability frameworks dedicated to dynamic environments exist (e.g., Keeney et al. 2006; Montanelli and Castano 2008), these frameworks focus on automatic semantic mapping to support run-time semantic interoperability. However, these approaches did not address the need for adapting the semantic interoperability solution to current changes of the ad hoc network.

As a result, existing frameworks for semantic interoperability are not comprehensive enough to ensure real time semantic interoperability in ad hoc networks of geospatial databases. We propose that a suitable framework can be developed based on a comparison between social networks, and ad hoc networks of geospatial databases and their users.

3.5 Social Network Theory: The Foundation of the Conceptual Framework

The concept of social networks was introduced by Barnes (1954) as a way to capture interactions between people. According to Contractor et al. (2006), a social network is a set of communicating agents and relationships among them. Agents can represent social groups, individuals, enterprises, authorities, or non-human agents such as databases or geospatial resources repertories (Contractor and Monge 2002); relationships include friendships and other social interactions, data transfers, business collaborations, etc. Notably, online communities are considered as social networks (Dholakia et al. 2004; Preece and Maloney-Krichmar 2005; Brown et al. 2007). Social networks support fast dissemination of information among their members; they also allow collaboration among agents having common interests but that are located in remote geographical regions, as well as the formation of "small worlds," where each node of the network has the ability to find short paths to other nodes without having a global knowledge of all existing connections (Duchon et al. 2006). Our framework is based on the idea that in real time semantically-interoperable networks, agents should be able to interact and exchange information in the same way as they do in social networks. Figure 3.1 indicates a list of social network properties, which are related to agents of the network, the network's organization, and relations and communication among agents.

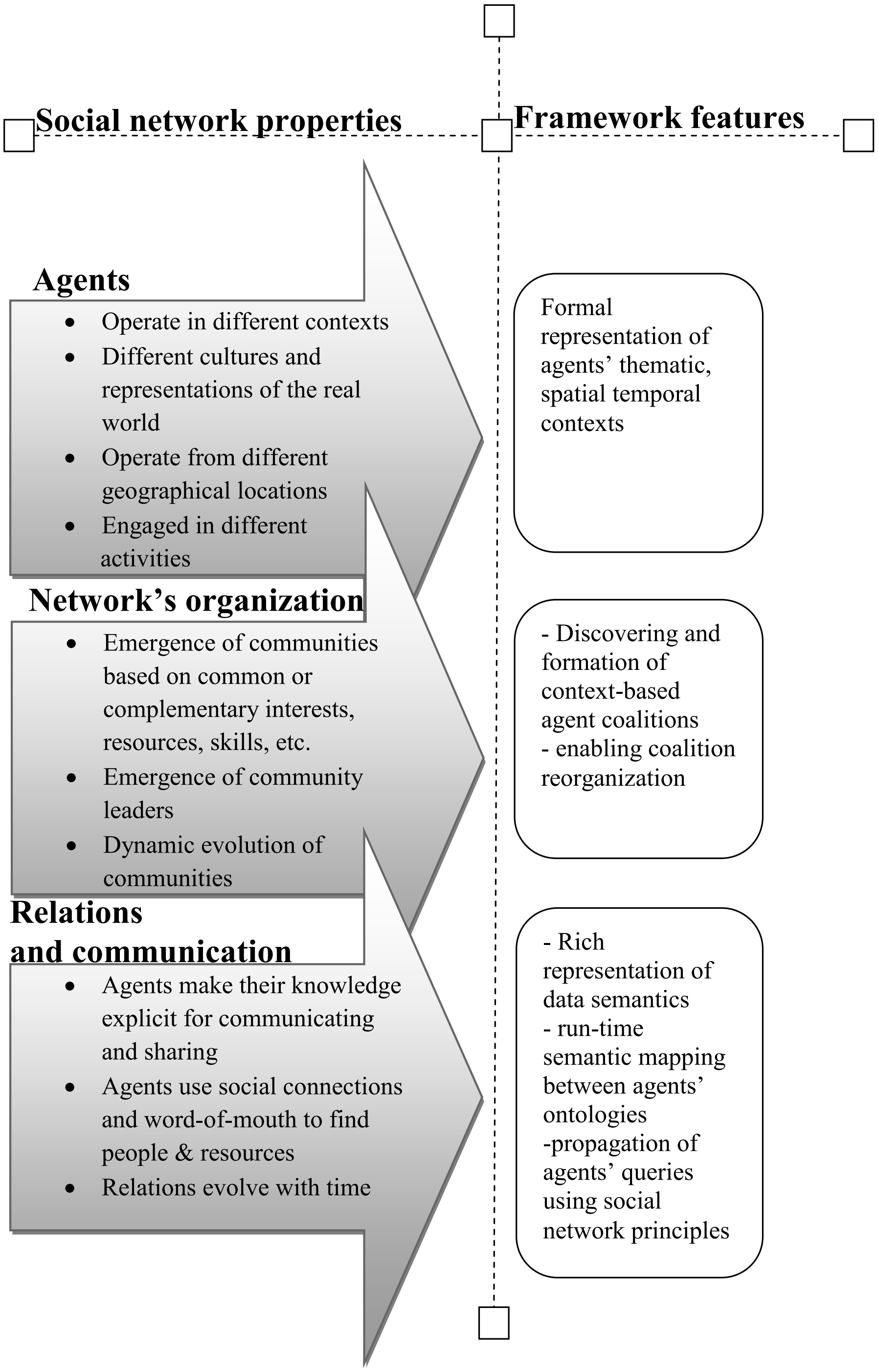

Social network properties
Framework features
Agents
Operate in different contexts
Different cultures and representations of the real world
Operate from different geographical locations
Engaged in different activities
Formal representation of agents' thematic, spatial temporal contexts
Network's organization
Emergence of communities based on common or complementary interests, resources, skills, etc.
Emergence of community leaders
Dynamic evolution of communities
- Discovering and formation of context-based agent coalitions
- enabling coalition reorganization
Relations and communication
Agents make their knowledge explicit for communicating and sharing
Agents use social connections and word-of-mouth to find people & resources
Relations evolve with time
- Rich representation of data semantics
- run-time semantic mapping between agents' ontologies
-propagation of agents' queries using social network principles

Figure 3.1 Desirable characteristics of a framework for real time semantic interoperability in ad hoc networks of geospatial databases inspired from social network properties

Those are desirable properties for a real time semantically-interoperable network; in Figure 3.1, they are translated into fundamental features for our conceptual framework:

- **Agents:** the framework should take into account the fact that agents operate in different contexts (culture, application domain, role, activities, geographical locations, etc.). Therefore, it requires explicit representation of the agents' contexts, in order to deal with these contexts, notably during the discovery and the formation of agent coalitions;
- **Network organization:** The framework should enable the creation of dynamic agent coalitions in the ad hoc network. Coalitions are groups of agents (with their database) that have similar contexts. To support the formation and management of coalitions in the ad hoc network, it is necessary to develop a coalition model and to formalize the different types of coalition reorganizations, such as coalition merging, division, dissolution, etc. In addition, to support the utilization of these coalitions in the semantic interoperability process, we should be able to define a global context of these coalitions, based on the context of their members.
- **Relations and communication:** in order to support meaningful communication and data exchange among agents, the framework should provide a rich representation of data semantics though ontologies. In order to find resources that match a user's query, the framework should enable the run-time establishment of semantic mappings between the query and the agents' ontologies. The framework should also include a model for the propagation of user's queries to relevant sources of the ad hoc network, using the social network properties highlighted in Figure 3.1.

3.6 A Conceptual Framework for Real Time Semantic Interoperability in Ad Hoc Networks of Geospatial Databases

The term "semantic interoperability" is linked to various notions, including the resolution of heterogeneities (Bishr 1998; Brodeur et al. 2003; Park and Ram 2004), communication (Brodeur et al. 2003; Carney et al. 2005), data sharing and exchange (Harvey et al. 1999; Rawat 2003; Brodeur et al. 2003; Carney et al. 2005), cooperation (Berners-Lee et al. 2001; Rawat 2003; Carney et al. 2005) and understanding of shared data (Brodeur et al. 2003;

Manso and Wachowicz 2009). Meanwhile, real-time systems are defined as reactive systems where accuracy of results and timely responses are critical (Lambert 2006). Kopetz (2011) also indicates that in a real-time computer system, "the correctness of the system behavior [i.e. the sequence of output in time of the system] depends not only on the logical results of the computations, but also on the physical time when these results are produced." Consequently, a real-time semantically-interoperable system is a system where the outputs of the components, such as responses to users' queries, are time-dependent. For example, the results of a query that was submitted in a relatively near past could be modified by the arrival of a new agent (with new source of data) in the ad hoc network. A real-time semantically-interoperable system should therefore be reactive to the events that modify the network: the adding of a new agent, but also the formation of a new coalition, the merging of existing coalitions, etc. This also means that users' queries are mapped on-demand to the agents' ontologies. We define a real-time semantically-interoperable system as "a reactive system that enables the agents to collaborate in order to retrieve and understand shared (geospatial) data they need at run-time." Based on this definition and on the social network properties, we propose our conceptual framework which is illustrated in Figure 3.2.

Let us assume the following situation. The ad hoc network is composed of multiple agents. Agents are entities that can act autonomously: they can enter or leave the network, join coalitions of other agents, and they can send queries to other agents. Each agent holds a geospatial database; the semantics of the data is formalized with an ontology. In addition, each agent holds an ontological description of the context of the geospatial database. In the following, the term "node" refers to the agent and the associated database and ontology.

First, in order to manage the large number of nodes, the ad hoc network is partitioned (step 1) into groups of nodes that have similar contexts: the coalitions. To obtain the context of a coalition, we merge the contexts and the coalition's members (step 2). For example, a coalition may hold data on the Quebec province's hydrographic network, for the purpose of water level measurement. Consider a user agent who is looking for flooding hazard areas. The user agent submits a query with the query concept "floodplains" (step 3). In addition to the query concept, the user agent specifies the context of its query, e.g., flooding risk assessment in Quebec City. When the user's query is sent in the network, it must be forwarded to the relevant nodes that are holding the requested data. Because the ad hoc network is decentralized, there is no central authority in charge of dispatching the query to the relevant nodes. Therefore, the query is propagated from node to node.

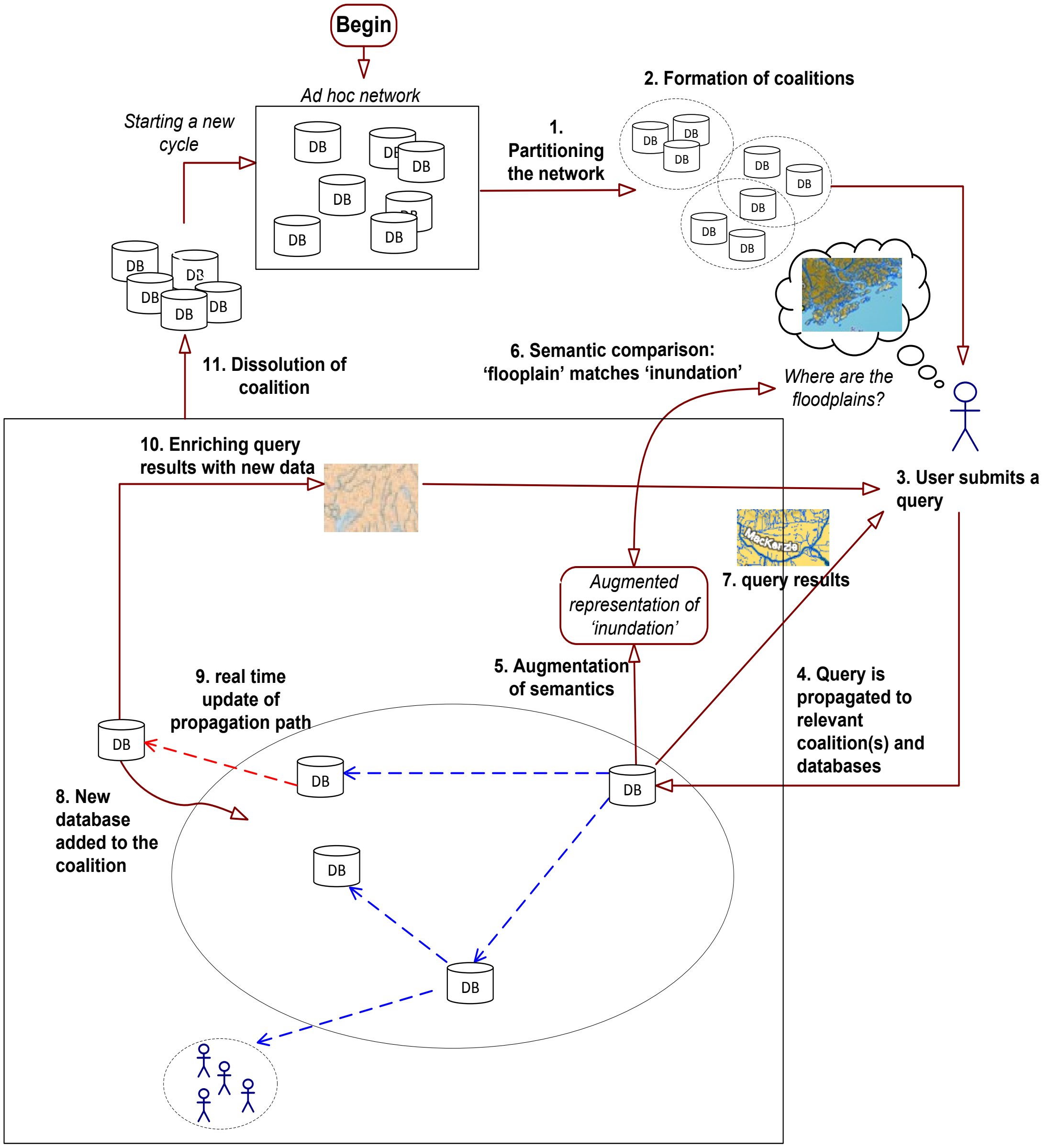

Begin
Ad hoc network
DB
Starting a new cycle
1. Partitioning the network
2. Formation of coalitions
3. User submits a query
Where are the floodplains?
4. Query is propagated to relevant coalition(s) and databases
5. Augmentation of semantics
Augmented representation of 'inundation'
6. Semantic comparison: 'flooplain' matches 'inundation'
7. query results
Mackenzie
8. New database added to the coalition
9. real time update of propagation path
10. Enriching query results with new data
11. Dissolution of coalition

Figure 3.2 Real Time Semantic Interoperability Conceptual Framework

The propagation of the query (step 4) is done in two steps. First, the query is propagated to the appropriate coalition(s); to do so, the query's context is compared to the context of available coalitions to find the relevant coalitions. Secondly, the query is propagated to the appropriate node(s), within the selected coalitions. The result of query propagation is a propagation graph, where the node(s) that are nearer to the graph's root are the more relevant to answer the query. As we move from the root node to the leaf nodes in the graph, the relevance of nodes with respect to the query decreases. The role of the propagation graph is therefore also to indicate which nodes are the more likely to contain the most relevant data with respect to the query. The propagation approach, which integrates social network properties mentioned on Figure 3.1, is presented in Chapter 6 of this thesis.

Once the relevant nodes are selected from the propagation graph, the requestor wants to know which concept of the targeted nodes'ontology best matches his query concept ("floodplains"). In order to enhance the matching process, the ontologies of targeted nodes are semantically augmented (step 5). The features that are added during the augmentation process help to clarify the contexts of the concepts and to improve the matching performance. The concepts that compose the augmented ontologies are called "multi-view augmented concepts" (MVACs). The MVAC model is presented in Section 3.6.2, and the MVAC generation approach is presented in Chapter 5 of this thesis. The query is compared to the concepts of the ontologies (step 6) with a qualitative semantic mapping system, the G-MAP, which is presented in Section 3.6.3, and its quantitative counterpart, the Sim-Net semantic similarity model, presented in Chapter 7 of this thesis. Based on the semantic matching results, the user can select the concept that best matches his query and retrieve the corresponding data (step 7).

If a new database is added to the network or joins the coalition (step 8), the propagation path is updated accordingly (step 9), as described in Chapter 6, and if a concept of the new node matches the user's query, the new data can be send to the requestor to enrich the existing query results with more recent data (step 10). Also, the coalitions can be dissolved and a new cycle to form new coalitions can be initiated (step 11).

In the rest of this paper, we present the semantic models that are needed to support this semantic interoperability process, including the coalition model, the augmented concept model, and the semantic mapping model that is adapted to this augmented concept model.

3.6.1 Geospatial Databases Coalition Model

Coalitions of geospatial databases enable the partitioning of the ad hoc network according to several context parameters; this partitioning supports the identification of relevant nodes during query propagation. For example, the network can be partitioned into coalitions based on the function of databases, e.g., environmental monitoring (Figure 3.3).

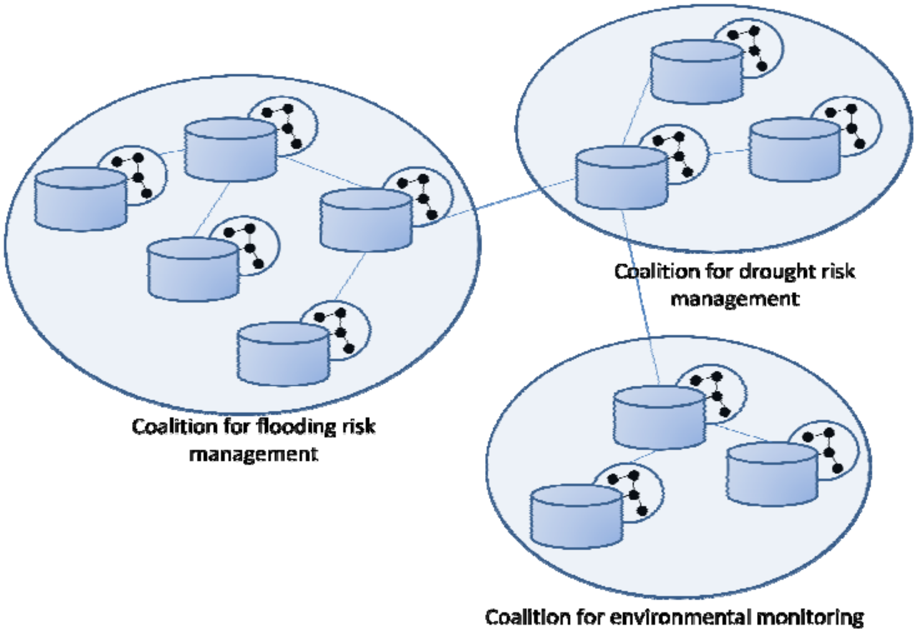

Coalition for drought risk management
Coalition for flooding risk management
Coalition for environmental monitoring

Figure 3.3 Example of geospatial databases coalitions

Coalitions represent the ability of members of a social network to auto-organize. The Coalition model is illustrated in Figure 3.4. The model's role is to capture the fundamental elements that will enable the formation and the description of coalitions. The network is composed of geospatial databases, which are semantically described by an ontology. Geospatial databases are also described by a context, which include the function of the database (i.e., the intended use of data), the domain described by the data(e.g., ecology, hydrography, transportation, etc.), the geographical location where the real world entities represented in the geospatial database are located (in other words, the spatial coverage of the data), and the temporal validity period during which the real world entities represented in the geospatial database exist The interests of users are expressed with the same parameters.

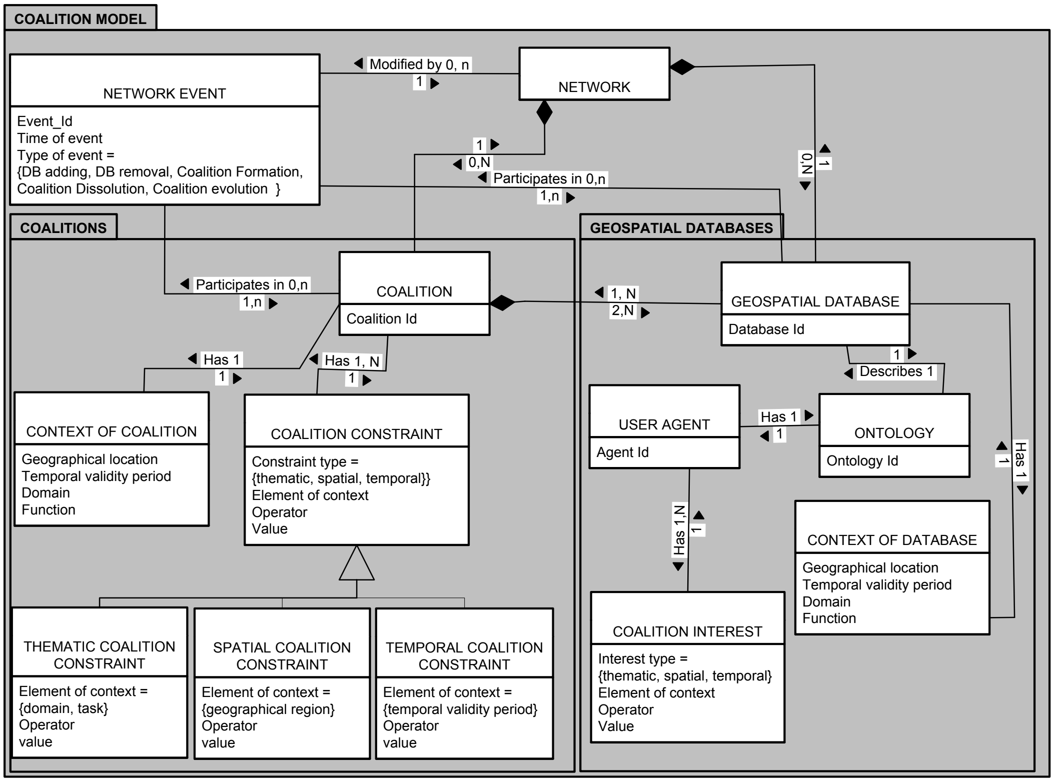


Figure 3.4 Network Coalition Model

Coalitions are groups of users (and their geospatial databases) that have a common goal, such as working for a common interest (land management, risk assessment, etc.), forming a group of expertise on the same domain (transport, weather, etc.), or combining complementary information on a same geographical location. Coalitions also have a context analogous to the context of databases, and constraints on the values of the context parameters that can be used to restrict the entry of additional geospatial databases in the coalition. Coalitions can be modified through "networks events," such as the addition or the removal of geospatial databases. Further details on context parameters, the generation of coalitions' context and network events are provided in Chapter 4.

3.6.2 New Formalization of Geospatial Concepts: The Multi-View Augmented Concept (MVAC) Model

The Multi-View Augmented Concept (MVAC) model is a new concept representation whose role is to enrich existing concepts of databases' ontologies with new features, but without using knowledge that is external to the geospatial database. The MVAC represents the ability of social network members to make their knowledge explicit as they communicate with others. The idea of the MVAC is to add two additional levels of knowledge to the original concept definition: a set of views, which represent the concept in different contexts, and a set of dependencies between features of the concept (Figure 3.5).

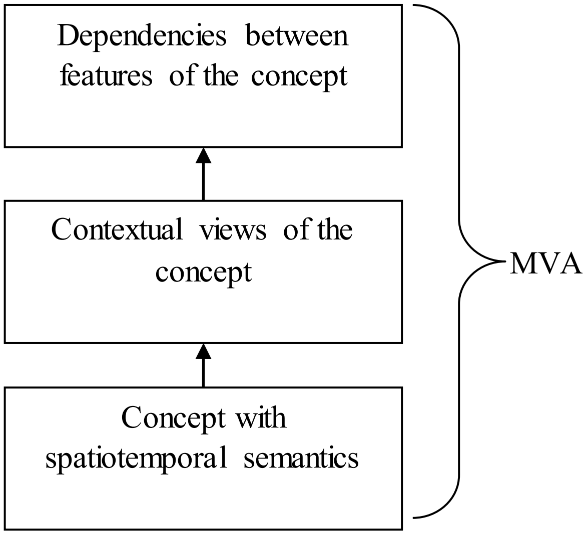

Dependencies between features of the concept
Contextual views of the concept
Concept with spatiotemporal semantics
MVA

Figure 3.5 Graphical representation of the MVAC model

Figure 3.6 provides the UML model of the MVAC. In the following sub-sections, we present the components of the MVAC: the views and contexts (3.6.2.1), the semantics of spatial and temporal properties (3.6.2.2), and the dependencies between the concept's features (3.6.2.3).

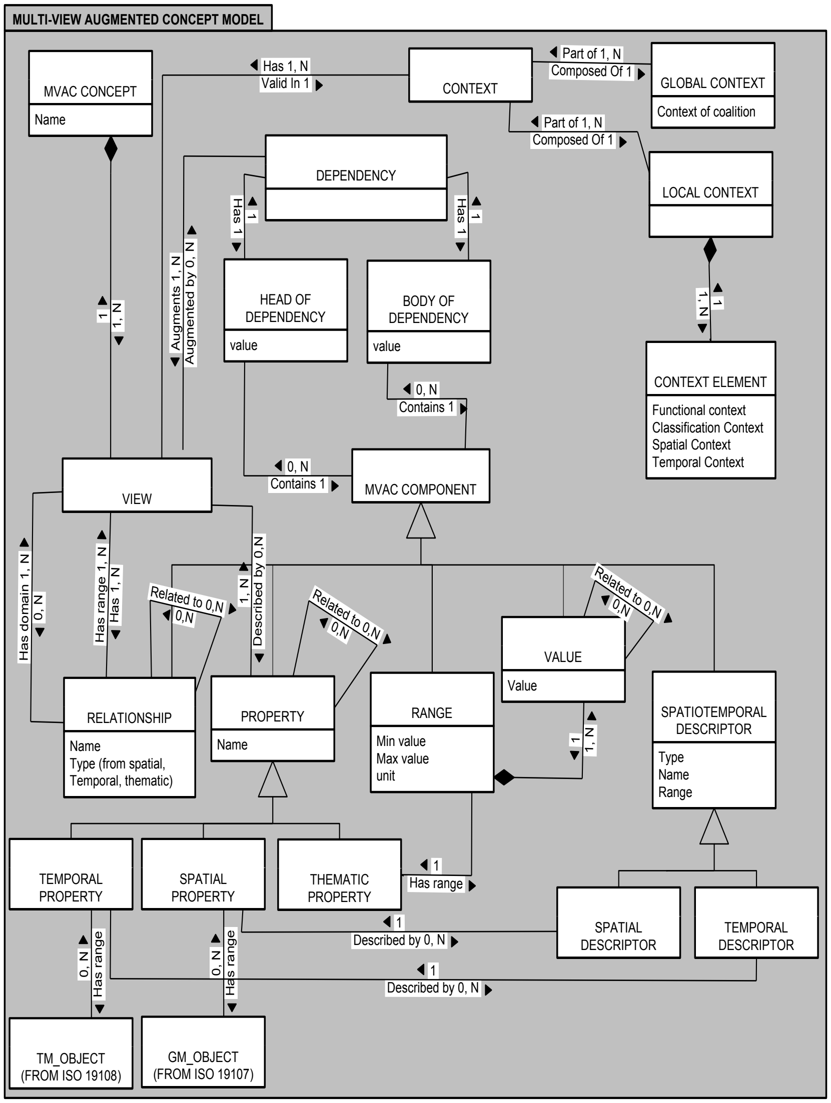


Figure 3.6 MVAC Model

3.6.2.1 Views and Context

The multi-view principle consists in considering that a concept may implicitly contain several sub-concepts (called views), where each view represents the concept in a given context. Then, as members of a social network, users of geospatial databases can adapt their communication strategy to the chosen context. In the geospatial domain, it is well known that a concept can be defined under different perspectives (Bédard and Bernier 2002; Parent et al. 2006). Views represent the same reality at different levels of detail, spatial resolution, with different geometries, or thematic features, etc. Views are also used in the ontology domain to represent ontology subsets (Bhatt et al. 2006; Wouters et al. 2008). These approaches implicitly recognize that a concept may be represented differently according to various contexts, but the context itself is not represented explicitly. In our approach, we distinguish between the context and the representation of the concept in this context (a view of the concept). We define a view as a selection of properties and relations (a subset of the concept's properties and relations) valid in a given context. Properties and relations are spatial, temporal thematic, or mixed (spatio-temporal, spatio-thematic, etc). As shown in Figure 3.6, each view is associated with a global context, which is the context of the coalition that the concept belongs to, and the local context of the view. In Chapter 5, we provide further details on the context representation that we have developed for MVACs.

3.6.2.2 Semantics of Spatial and Temporal Properties

Each view of the MVAC is associated with a description of the semantics of its spatial and temporal properties; the semantics of spatial and temporal properties is expressed through a set of basic elements, called spatial and temporal descriptors. The semantics of a spatial property (e.g., the spatial property "geometry" of a concept "house") is described by the specific spatial entity that the geometry represents in reality (e.g., the spatial entity can be the roof or the basement of the house) (Figure 3.7).

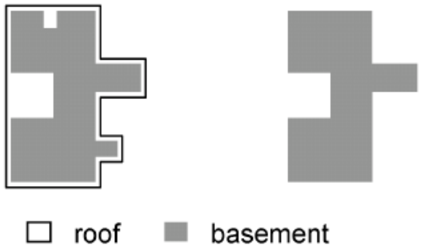


Figure 3.7 Example of different spatial semantics for "house" (from Brodeur 2004)

In geospatial database specifications, the spatial entity is often described as a part of a thing, for instance "center of, axis of, edge of, contour of, on top of …" (Gesbert 2005). In addition, spatial descriptors include the characteristics related to geometry: the shape, spatial attributes (area, length, etc.). The semantics of a temporal property is described by the temporal entity that it represents in reality. The temporal entity is an occurrent: a process, event, activity or change that unfolds itself through a period of time (Grenon and Smith 2004). Temporal descriptors include temporal attributes, such as duration and frequency. If the temporal entity is a period, it is described by a *start* and an *end* that give the semantics of the period's boundaries. If the temporal entity is an instant, it is described by an *event* that gives the semantics of this point in time. For example, the duration of a storm may be defined by the start and end events *formation of clouds* and *end of precipitation*.

3.6.2.3 Dependencies between Features

The features of a concept are the spatial, temporal and thematic properties; the spatial, temporal and thematic relations; and the spatial and temporal descriptors. We represent dependencies between features in order to augment the expressivity of concepts. Dependencies express a constraint where the value of a first feature is constrained by thevalue of a second feature. Dependencies are considered as valuable properties of expressive ontologies (Curé and Jeansoulin 2007). For example, *temperature* depends on *altitude*. Dependencies may be valid in one, some or all of the views of a concept. Dependencies are formalized as rules, in the form: head → body. The body is the consequence of the head. Body and head are composed of dependency elements, which involve concepts or views, as well as the concept's features. Table 3.1 gives the set of possible dependency elements we propose. They complement the set of elements defined for rules in Brockmans and Haase (2006). The variable x and y represent instances of concepts, while the variable z represents a value of a property or descriptor.

Table 3.1 Dependency Elements

Dependency Element	**Form**	**Meaning**
View dependency element	v(x)	x is an instance of view v
Concept dependency element	c(x)	x is an instance of concept c
Thematic property dependency element	$p_{th}(x, z)$	The value of property p_{th} for x is z, where z is a thematic value
Spatial property dependency element	$p_s(x, z)$	The value of property p_s for x is z, where z is a geometry type
Temporal property dependency element	$p_t(x, z)$	The value of property p_t for x is z, where z is a temporal type

Thematic relation dependency element	$r_{th}(x, y)$	x and y are linked by relation r_{th}
Spatial relation dependency element	$r_s(x, y)$	x and y are linked by relation r_s, with x and y both having geometrical extent
Temporal relation dependency element	$r_t(x, z)$	x and y are linked by relation r_t, with x and y both having temporal extent
Spatial descriptor dependency element	$d_s(x, z)$	The value of descriptor d_s for x is z, where z is a spatial descriptor value
Temporal descriptor dependency element	$d_t(x, z)$	The value of descriptor d_t for x is z, where z is a temporal descriptor value

In Chapter 5, we provide a method for generating MVAC concepts. However, we note that according to related work, existing semantic mapping approaches lack criteria that are required for comparing all elements of the MVAC concepts, notably, the ability to take into account semantics of spatial and temporal properties and to compare dependencies. This reduces the ability of these approaches to resolve semantic heterogeneities. This is why in the next section, we propose a semantic mapping model that is adapted to the MVAC model.

3.6.3 G-MAP Augmented Semantic Mapping Model for Geospatial Databases

In the proposed framework, users share explicit representations of the semantic relations, or *semantic mappings*, which link their geospatial concepts. Semantic mappings are used to translate knowledge between geospatial databases. In social networks, small groups of people can communicate based on local interaction; they do not need a global agreement over the whole network. Similarly, semantic mappings are local bridges between geospatial database users. Since there is no a priori semantic agreement between geospatial databases users, the semantic mapping process is automatic. The G-MAP semantic mapping model introduces a new gradual semantic mapping process; G-MAP also has the ability to compareMVAC concepts. G-MAP matches the most basic elements of the MVAC and reuses the results to match complex elements of the MVAC. The contribution of G-MAP with respect to existing matching systems is its ability to consider a higher degree of complexity in the concept's representation, with well-defined spatiotemporal features and new matching criteria, which allows discovering implicit mappings. In addition, in previous research, we have demonstrated that G-MAP can be applied to enable and improve semantic interoperability of geospatial web services (Bakillah and Mostafavi 2010).

For the purpose of the G-MAP algorithm, MVAC elements are classified into *Basic MVAC Element*, which are names of all features, and *Complex MVAC Element*, which are properties, relations, context, etc. The G-MAP gradual process is composed of the sub-processes illustrated in Figure 3.8.

The Semantic Matching of *Basic MVAC Element* is performed by the *Basic MVAC Element Matcher*, and the semantic matching of *Complex MVAC Element*, of views and of MVAC by the *G-MAP Mapping Inference Engine*. The augmented semantic matching process is performed by the *Augmented Mapping Inference Engine*.

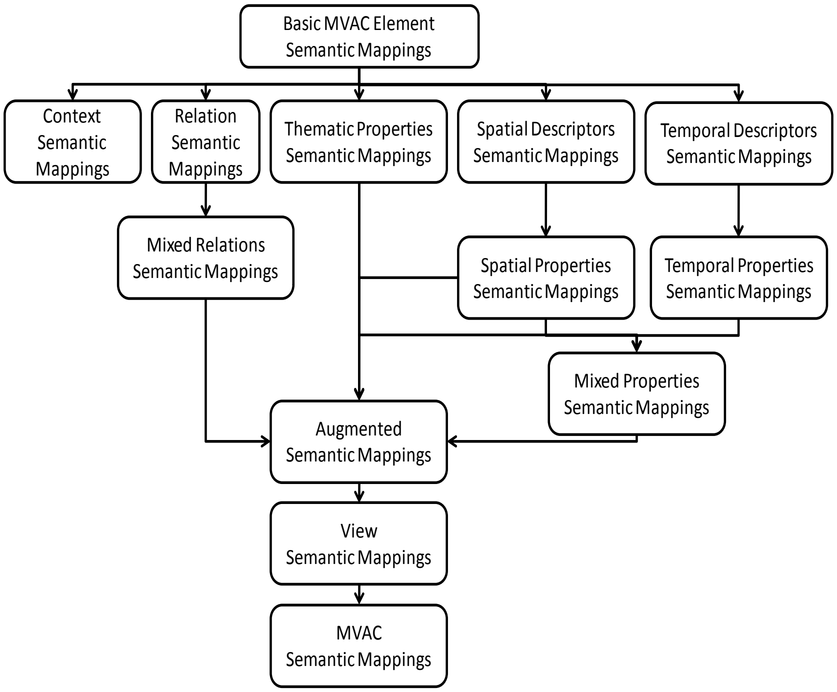

Basic MVAC Element Semantic Mappings
Context Semantic Mappings
Relation Semantic Mappings
Thematic Properties Semantic Mappings
Spatial Descriptors Semantic Mappings
Temporal Descriptors Semantic Mappings
Mixed Relations Semantic Mappings
Spatial Properties Semantic Mappings
Temporal Properties Semantic Mappings
Mixed Properties Semantic Mappings
Augmented Semantic Mappings
View Semantic Mappings
MVAC Semantic Mappings

Figure 3.8 Gradual matching process of the G-MAP

3.6.3.1 Basic MVAC Element Matcher

The role of the Basic MVAC Element Matcher is to find semantic relations between basic elements of different MVAC. We use the expression "basic element" to refer to any label that is part of the definition of a concept: the name of the concept, names of properties, of relations, of descriptors, of contexts, and names of properties and descriptors' values. Many semantic mapping approaches also include an equivalent matching phase, often called pre-processing, for example in S-Match (Giunchiglia et al. 2004). In addition, many of those approaches use external resources (such as WordNet) to find the lexical relations between terms. The contribution of the Basic MVAC Element Matcher with respect to those approaches is the specific processing of basic elements used to denote spatial, temporal and spatiotemporal features (properties, relations or descriptors). The Basic MVAC Element Matcher uses external resources (more specifically, global ontologies) that are appropriate for spatial and temporal features. Furthermore, it uses the additional knowledge provided by spatial and temporal descriptors to refine the basic matching of names of spatial and temporal properties and relations. Figure 3.9 illustrates the functional architecture of the Basic MVAC Element Matcher.

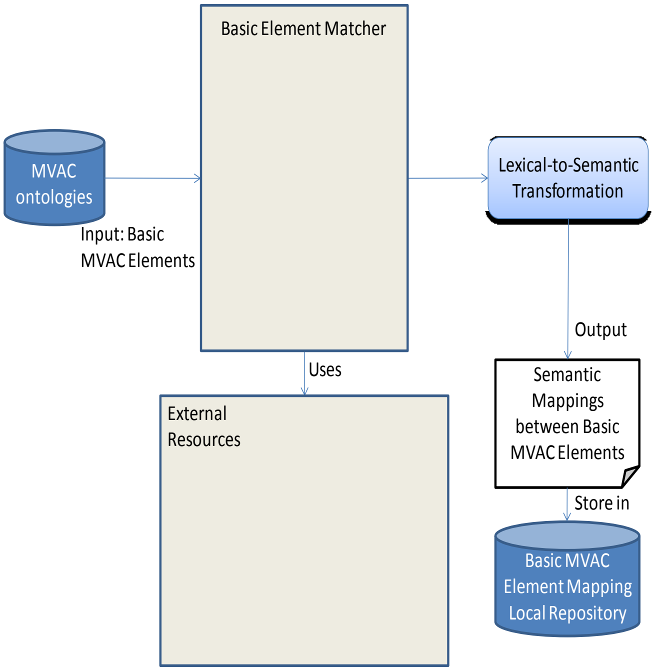

Basic Element Matcher
MVAC ontologies
Input: Basic MVAC Elements
Lexical-to-Semantic Transformation
Output
Uses
External Resources
Semantic Mappings between Basic MVAC Elements
Store in
Basic MVAC Element Mapping Local Repository

Figure 3.9 Basic MVAC Element Matcher's functional architecture

The basic matching process is performed in two main steps. First, it finds the relation between each pair of basic elements that denote the same type of features (thematic, spatial, or temporal) using the appropriate matcher. The Basic MVAC Element Matcher is able to identify the appropriate matcher since in the MVAC every feature is “tagged” with its type. The relation between basic elements is identified with the help of the appropriate external resource. The nature of this relation depends on the external resource. For example, to compare thematic basic features, we use WordNet, which issues lexical relations (synonym, hypernymy, hyponym). However, some external resources issue relations such as “specialization of” and “generalization of”. An example of this type of external resource is the Open Cyc Ontology of Spatial Relations. In the second step, the Basic MVAC Element Matcher transforms the relation issued in the first step into one of the following set-based semantic relations: equivalent, includes, included in, disjoint. The transformation is performed to make the output of the first step uniform, whatever the external resource being used. The resulting relations have to be uniform since they will be reused by the next component of G-MAP, namely the G-MAP Complex Mapping Inference Engine. Table 3.2 shows the transformation for lexical relations issued by WordNet between thematic basic elements. The t(x) notation indicates that the variable x is a basic element.

Table 3.2 Transformation rules

Transformation rules
$t(x) \wedge t(y) \wedge \text{synonym}(x, y) \Rightarrow \text{equivalent}(x, y)$
$t(x) \wedge t(y) \wedge \text{hypernym}(x, y) \Rightarrow \text{includes}(x, y)$
$t(x) \wedge t(y) \wedge \text{hyponym}(x, y) \Rightarrow \text{included in}(x, y)$
$t(x) \wedge t(y) \wedge \text{disjoint}(x, y) \Rightarrow \text{disjoint}(x, y)$

This transformation is based on the one proposed by Serafini et al. (2003). It relies on the fact that lexical relations have set-theoretic implications. If we consider that a term has an extensional definition (that is, the set of real world objects that it represents), then a first term x (ex: waterbody) which is a hypernym of a second term y (ex: lake) includes y, since all objects that y represents can also be classified under x. Retrieving lexical relations using WordNet is common in the existing literature (e.g. Serafini et al. 2003; Giunchiglia et al. 2004). However, in the next sections, we explain the specific process for spatial and temporal features.

3.6.3.1.1 The Case of Spatial Features and Temporal Features

Many semantic mapping approaches have focused on the thematic aspects of concepts, leaving aside the complexities of their spatial and temporal aspects. While some semantic approaches are dedicated to geospatial databases, such as Hess et al. (2007), they still consider spatial properties as point, line or polygon primitives only, but do not take into account more complex spatiotemporal semantics such as in the MVAC model, where spatial and temporal properties are further enriched with descriptors, and instances of concepts can be linked with various spatiotemporal relations (such as “close to,” “above,” etc.). In this Section, we present a novel approach to deal with spatiotemporal semantics during the semantic mapping process.

Concepts in geospatial databases are described with spatial features that are specific to the geospatial domain, especially spatial relations of topology, proximity and orientation, attributes of spatial features (length, area, width) and geometrical shapes. For example, some spatial relations have been formalized and given a well-defined meaning (Egenhofer’s topological relations, 1993); however, Schwering (2006) notes that many other spatial relations can relate instances of concepts, such as “above,” “below,” “across,” etc. General external resources such as WordNet are not suitable to retrieve semantic relations between names of spatial features. For example, WordNet does contain the term “above”, but does not distinguish between “above-directly”, “above-touching”, or “above-higher” (Figure 3.10).

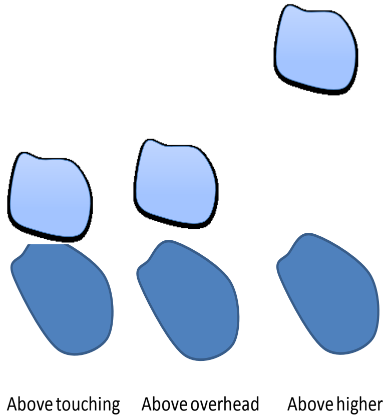


Figure 3.10 Different meanings of the spatial relation “above”

For example, an antenna can be "above-touching" a building; a road sign can be "above-overhead" a street; while hazard to air navigation can be "above-higher" the ground. The case of temporal features raises a similar concern. While, for example, temporal relations were formally defined by Allen (1983), many other temporal concepts are employed to describe geospatial concepts' temporal features. For example, Figure 3.11 shows different meanings of the temporal relation "temporal bound intersect," which means that the boundary of a first time interval intersects a second time interval. In Figure 3.11, the hierarchical relations are generalization/specialization relations. When the relation "temporal bound intersect" is verified between two time intervals, it could mean that the time intervals are intersecting ("temporally intersect"); that the end of the first time interval happens during the second time interval ("ends during"); that the beginning of the first time interval happens during the second time interval ("starts during"); and so on. These temporal relations can be used to describe relations between events.

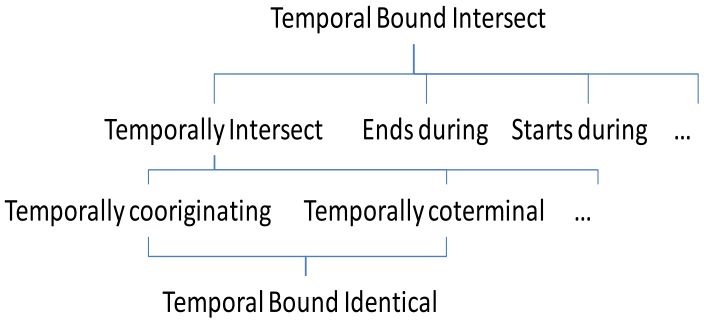


Figure 3.11 Different meanings of the temporal predicate "temporal bound intersect"

Therefore, we must rely on comprehensive ontologies of spatial and temporal concepts. Ontologies of spatial (or temporal) concepts describe spatial (or temporal) concepts in general, such as shapes, spatial relations, and temporal relations, regardless of the application domain. For the purpose of this approach, we used the OpenCyc spatial properties and relations ontology, as well as the temporal component of OpenCyc. Cyc is an artificial intelligence (AI) project that aims at creating a comprehensive ontology and knowledge base of common sense knowledge to support human-like reasoning of AI applications[3]. OpenCyc Spatial Ontology contains concepts related to shape attributes, direction and orientation, relative position of objects, and mereological relations, to name a few. As for its temporal component, it describes time and date concepts, relations between temporal objects (instants and periods), time measurement units, and properties of temporal objects. The concepts in Open Cyc are related with specialisation relations (the inverse of the generalisation relation). However, we note that the principle of our approach is independent of the chosen external resource.

The Basic Spatial and Temporal Element Matchers' role is to find the semantic relations between the pairs of spatial (respectively temporal) features of different concepts, with the help of the Spatial (respectively Temporal) Ontology; in this case, we use the Open Cyc Ontologies. Note that these features include the "spatial component" and "temporal component" of spatiotemporal features, which are treated separately and combined later. The idea of the Basic Spatial and Temporal Element Matchers is to perform several progressive steps to find, in the Spatial (Temporal) Ontology, the most specific concepts (i.e., the lowest possible concepts in the hierarchy) that match each of the spatial or temporal basic elements being compared. This most specific concept represents the meaning of the basic element. The more precisely we can identify this specific concept, the more the matching will be accurate. Then, the relation between those basic elements is retrieved by identifying in the Spatial (Temporal) Ontology the relation between those most specific concepts. Figure 3.12 illustrates this principle for the spatial case. The semantic relation between two spatial relations O1:r and O2:r that belong to different ontologies is derived from the relation that holds between the most specific concepts of the Spatial Ontology they are referring to.

[3] http://www.opencyc.org/

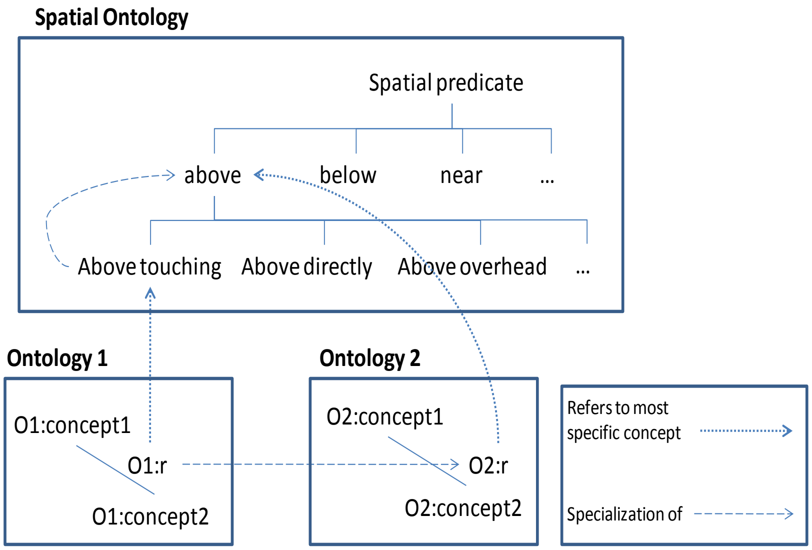


Figure 3.12 Deriving the relation between spatial features' names from the relation between the most specific concepts they refer to in the Spatial Ontology

Figure 3.13 illustrates the procedure that we have developed to retrieve the most specific spatial (temporal) concept of the Spatial (Temporal) Ontology that matches a spatial (temporal) feature's name. The name of a spatial (temporal) feature is hereafter referred to as a basic element *El*. The Spatial or Temporal Ontology is hereafter referred to as the Global Ontology. The proposed procedure is divided in two sub-procedures:

- the First Level Match Procedure: the goal of this procedure is to find, in the Global Ontology, the concept *C* that is closer to the element *El*.

- the Refine First Match Procedure. The goal of this procedure is to take the concept *C* of the Global Ontology that was identified in the First Level Match Procedure, and try to refine this match by determining if the (spatial and temporal) descriptors of *El* could help to precise the meaning of *El*.

First Level Match Procedure: The first step of this procedure is to verify if the Global Ontology contains *El*. If a perfect match is found (i.e., the Global Ontology contains a concept whose name exactly matches the term *El*), the resulting concept *C* is stored and the First Level Match Procedure is completed. However, if no such exact match is found, we try to determine if one of the WordNet synonyms of El could be an exact match of a concept from the Global Ontology.

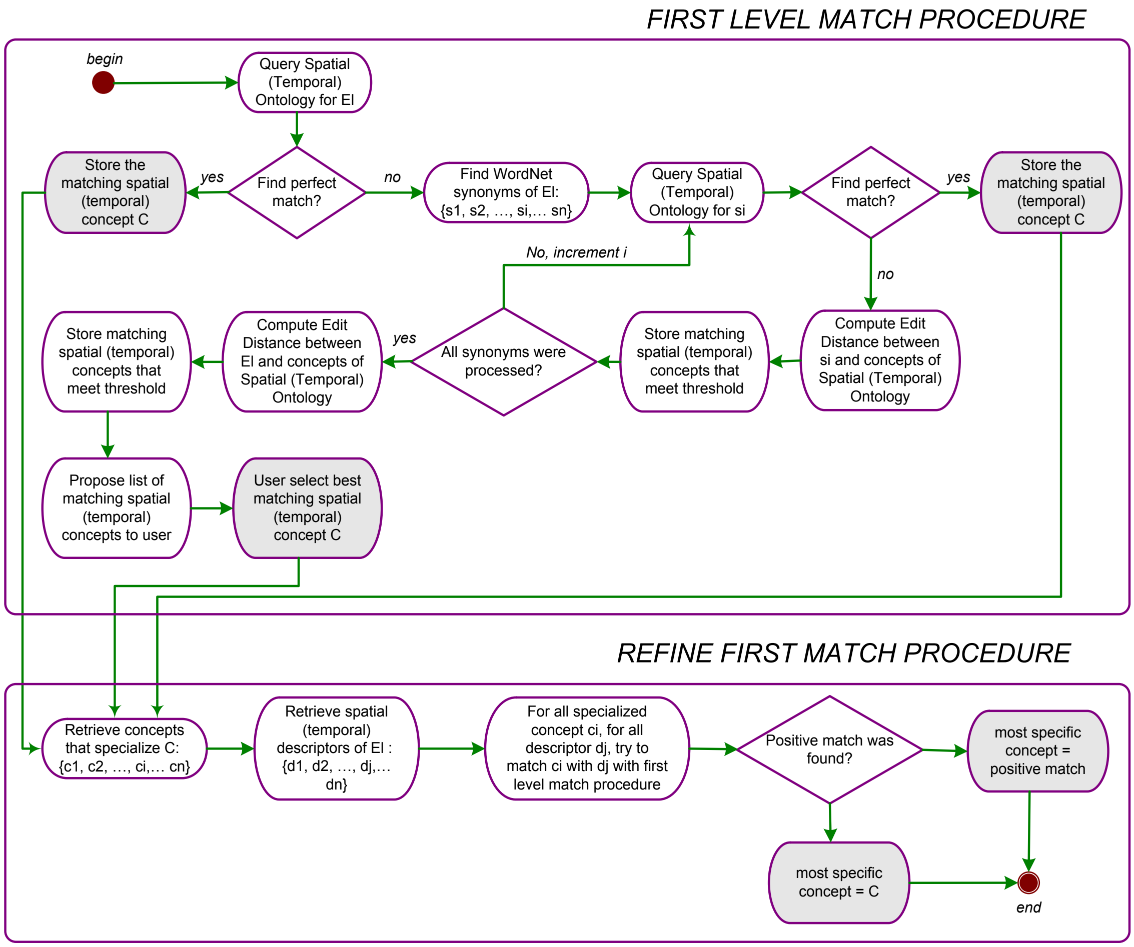

FIRST LEVEL MATCH PROCEDURE
begin
Query Spatial (Temporal) Ontology for El
Store the matching spatial (temporal) concept C
yes
Find perfect match?
no
Find WordNet synonyms of El: {s1, s2, …, si,… sn}
Query Spatial (Temporal) Ontology for si
Find perfect match?
yes
Store the matching spatial (temporal) concept C
No, increment i
no
Store matching spatial (temporal) concepts that meet threshold
Compute Edit Distance between El and concepts of Spatial (Temporal) Ontology
yes
All synonyms were processed?
Store matching spatial (temporal) concepts that meet threshold
Compute Edit Distance between si and concepts of Spatial (Temporal) Ontology
Propose list of matching spatial (temporal) concepts to user
User select best matching spatial (temporal) concept C
REFINE FIRST MATCH PROCEDURE
Retrieve concepts that specialize C: {c1, c2, …, ci,… cn}
Retrieve spatial (temporal) descriptors of El : {d1, d2, …, dj,… dn}
For all specialized concept ci, for all descriptor dj, try to match ci with dj with first level match procedure
Positive match was found?
most specific concept = positive match
most specific concept = C
end

Figure 3.13 Two-level procedure for retrieving the most specific concept that describe the meaning of spatial and temporal features

If an exact match is found in this second attempt, the resulting concept *C* is stored and the First Level Match Procedure is completed. If again, no exact match is found for a synonym of *El*, we try to find a partial match between this synonym and the concepts of the Global Ontology. For this, we employ the Edit Distance, which is defined as the minimum number of character insertions, deletions or changes necessary to turn one string into another (Cormode and Muthukrishnan 2007). The concepts of the Global Ontology for which the Edit Distance with a synonym of *El* are lower than a given maximal confidence threshold are stored as potential matches. A similar procedure is conducted with *El*, that is, the Edit distance is measured between *El* and the concepts of the Global Ontology, and the concepts of the Global Ontology that are below the confidence threshold are stored as potential matches. Finally, potential matches are displayed to the user that must select the best match; this completes the First Level Match Procedure.

Refine First Match Procedure: This procedure tries to match the descriptors of *El* with the concepts that specialize *C* in the Global Ontology, using the same steps that compose the First Level Match Procedure. If a positive match is found, this positive match is selected as the most specific concept of the Global Ontology that matches *El*. If no positive match is found, it means that *C* is the most specific concept of the Global Ontology that matches *El*.

When the most specific concepts, say C1 and C2, to which two basic elements El1 and El2 refer have been found following the above procedure, the final step is to retrieve the relation between them. In OpenCyc spatial and temporal Ontologies, the concepts are related by “specialization of” and its inverse “generalization of” relations. The “equivalent” relation is produced when C1 = C2, and disjoint when no relation can be inferred between them. The relation can be retrieved directly (when C1 and C2 are directly related in the Global Ontology) or using transitive relations formalized in the Global Ontology. A transitive relation is a relation R such that for any concept C1, C2 and C3, if R(C1, C2) and R(C2, C3), then R(C1, C3). Examples of transitive relations are “includes”, “located in”. The results of the Basic MVAC Element Matcher are stored in the Basic MVAC Element Mapping Local Repository to be reused in the next process.

3.6.3.2 G-MAP Complex Mapping Inference Engine

The role of the G-MAP Complex Mapping Inference Engine is to compute semantic relations between Complex MVAC Elements, between views and between MVACs. The problem of computing those semantic relations is formulated as the problem of verifying a set of logical mapping rules, which express the condition for a semantic relation between two elements to be true. The principle underlying the mapping rules we have developed is based on set theory. We assume that a concept's definition is of the "intentional" type: an intentional definition consists of a description of the characteristics that a real world entity should have to be considered as an instance of the concept (Castillo et al. 2003). Therefore, the meaning of the semantic expression: "two concepts are overlapping" is that those two concepts can have a common set of instances. This is a very strict condition, since two concepts that share several properties but for which one of the common properties have disjoint ranges of values would be considered as non-overlapping (disjoint). For example, if a concept C1 is "a river whose width is between 7 and 10 m" and a concept C2 is "a river whose width is between 11 and 20 m", C1 and C2 would be disjoint. While it is true that no river can be an instance of both C1 and C2, in several cases, these concepts would be considered as close enough in meaning, especially if the user does not care about a spatial constraint on the width of the rivers. Consequently, we consider that concepts which cannot share any instance but that have at least one non-disjoint feature are "weakly ovelapping", while pairs of concepts that verify the strict condition are "strongly overlapping". Table 3.3 shows the meaning of each possible relation that the G-MAP Complex Mapping Inference Engine can infer between concepts or between features. Note that relations between features are computed by considering features as concepts being defined with a single feature.

Table 3.3 Meaning of semantic relations that indicate how mapping rules are defined

Equivalent(C1, C2)	Includes(C1, C2)	Included in(C1, C2)	Strong overlap(C1, C2)	Weak overlap(C1, C2)	Disjoint(C1, C2)
All instances of C1 are instances of C2, and all instances of C2 are instances of C1	All instances of C2 are instances of C1, but not all instances of C1 are instances of C2	All instances of C1 are instances of C2, but not all instances of C2 are instances of C1	Some but not all instances of C2 are instances of C1, and some but not all instances of C1 are instances of C2	No instances of C2 are instances of C1, and no instances of C1 are instances of C2, but C1 and C2 have some common features	No instances of C2 are instances of C1, and no instances of C1 are instances of C2

The computation of semantic relations between Complex MVAC elements, views and MVACs is based on the principle of rule-based inference systems. A rule-based inference system takes as input facts from a fact base, and produces new facts inferred by rules being stored in a knowledge base in a recursive manner. Our *G-MAP Complex Mapping Inference Engine* uses G-MAP semantic mapping rules which each consists of a mapping rule antecedent and a mapping rule consequent (Figure 3.14):

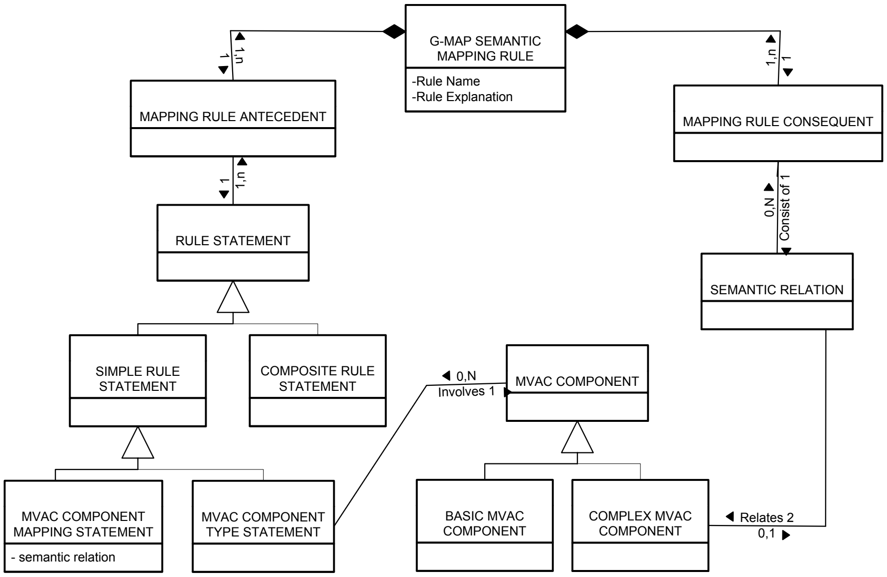


Figure 3.14 G-MAP mapping rule model

- The mapping rule antecedent is a conjunction of rule statements that must be verified. Rule statements can be simple (composed of a single statement) or composite (composed of several statements related with logical Boolean expressions *and* ($\wedge$), and *or* ($\vee$)). There are two different types of simple rule statements:
  - *MVAC element type statement* is a statement about the nature of a MVAC element x; for example, the statement $p(x)$ indicates that x is a property.
  - *MVAC element mapping statement* is the affirmation of a semantic relation between two MVAC elements, which are of the form *relation*(x, y), where x, y are MVAC elements and may be derived from the Basic MVAC Element Matcher or the Complex MVAC Element Matcher.

- The mapping rule consequent is the consequence of the antecedent. It is a semantic relation that holds between two complex MVAC elements that are part of the definition of two different MVACs.

The basic G-MAP Semantic Mapping Rules we have developed for the different types of MVAC complex elements (properties, relations, context, views and MVAC) are listed in Table 3.4. The principle for computing semantic relations between mixed features (e.g. spatiotemporal properties, spatiotemporal relations, etc.) is that mixed features are considered as concepts defined by a conjunction of two properties, for example: spatiotemporal_property(x) = spatial_property(y) ∧ temporal_property(z).

P(x) means that x is a property; name (x, np1) means that np1 is the name of x, and range (x, rp1) means that rp1 is the range of x. Spatial_descriptors (x, sd1) means that sd1 is a spatial descriptor of x. Equivalent (np1, np2) means that np1 is equivalent to np2, and similarly for the other semantic relations. The ¬ sign indicates negation; ∧ means "and," while ∨ means "or."

Table 3.4 G-MAP Semantic Mapping Rules

G-MAP Semantic Mapping Rules
Mapping of spatial properties :
p(x) ∧ p(y) ∧ name (x, np1) ∧ name (y, np2) ∧ range (x, rp1) ∧ range (y, rp2) ∧ spatial_descriptors (x, sd1) ∧ spatial_descriptors (y, sd2) ∧ equivalent (np1, np2) ∧ equivalent (rp1, rp2) ∧ equivalent (sd1, sd2) ⇒ equivalent (x, y)
p(x) ∧ p(y) ∧ name (x, np1) ∧ name (y, np2) ∧ range (x, rp1) ∧ range (y, rp2) ∧ spatial_descriptors (x, sd1) ∧ spatial_descriptors (y, sd2) ∧ [¬disjoint(np1, np2) ∨ overlap (rp1, rp2) ∨ overlap (sd1, sd2)] ⇒ weak_overlap (x, y)
p(x) ∧ p(y) ∧ name (x, np1) ∧ name (y, np2) ∧ range (x, rp1) ∧ range (y, rp2) ∧ spatial_descriptors (x, sd1) ∧ spatial_descriptors (y, sd2) ∧ [¬disjoint (np1, np2) ∧ overlap (rp1, rp2) ∧ overlap (sd1, sd2)] ⇒ strong_overlap (x, y)
p(x) ∧ p(y) ∧ name (x, np1) ∧ name (y, np2) ∧ range (x, rp1) ∧ range (y, rp2) ∧ spatial_descriptors (x, sd1) ∧ spatial_descriptors (y, sd2) ∧ [includes (np1, np2) ∨ equivalent (np1, np2)] ∧ [includes (rp1, rp2) ∨ equivalent (rp1, rp2)] ∧ [includes (sd1, sd2) ∨ equivalent (sd1, sd2)] ⇒ includes (x, y)
p(x) ∧ p(y) ∧ name (x, np1) ∧ name (y, np2) ∧ range (x, rp1) ∧ range (y, rp2) ∧ spatial_descriptors (x, sd1) ∧ spatial_descriptors (y, sd2) ∧ [included in (np1, np2) ∨ equivalent (np1, np2)] ∧ [included in (rp1, rp2) ∨ equivalent (rp1, rp2)] ∧ [included in (sd1, sd2) ∨ equivalent (sd1, sd2)] ⇒ included in (x, y)
p(x) ∧ p(y) ∧ name (x, np1) ∧ name (y, np2) ∧ range (x, rp1) ∧ range (y, rp2) ∧ spatial_descriptors (x, sd1) ∧ spatial_descriptors (y, sd2) ∧ [disjoint (np1, np2) ∨ disjoint (rp1, rp2) ∨ disjoint (sd1, sd2)] ⇒ disjoint (x, y)

Mapping of temporal properties :

p(x) $\wedge$ p(y) $\wedge$ name (x, np1) $\wedge$ name (y, np2) $\wedge$ range (x, rp1) $\wedge$ range (y, rp2) $\wedge$ temporal_descriptors (x, td1) $\wedge$ temporal _descriptors (y, td2) $\wedge$ equivalent (np1, np2) $\wedge$ equivalent (rp1, rp2) $\wedge$ equivalent (td1, td2) $\Rightarrow$ equivalent (x, y)

p(x) $\wedge$ p(y) $\wedge$ name (x, np1) $\wedge$ name (y, np2) $\wedge$ range (x, rp1) $\wedge$ range (y, rp2) $\wedge$ temporal_descriptors (x, td1) $\wedge$ temporal_descriptors (y, td2) $\wedge$ [$\neg$disjoint (np1, np2) $\vee$ overlap (rp1, rp2) $\vee$ overlap (td1, td2)] $\Rightarrow$ weak_overlap (x, y)

p(x) $\wedge$ p(y) $\wedge$ name (x, np1) $\wedge$ name (y, np2) $\wedge$ range (x, rp1) $\wedge$ range (y, rp2) $\wedge$ temporal_descriptors (x, td1) $\wedge$ temporal_descriptors (y, td2) $\wedge$ [$\neg$disjoint (np1, np2) $\wedge$ overlap (rp1, rp2) $\wedge$ overlap (td1, td2)] $\Rightarrow$ strong_overlap (x, y)

p(x) $\wedge$ p(y) $\wedge$ name (x, np1) $\wedge$ name (y, np2) $\wedge$ range (x, rp1) $\wedge$ range (y, rp2) $\wedge$ temporal_descriptors (x, td1) $\wedge$ temporal_descriptors (y, td2) $\wedge$ [includes (np1, np2) $\vee$ equivalent (np1, np2)] $\wedge$ [includes (rp1, rp2) $\vee$ equivalent (rp1, rp2)] $\wedge$ [includes (td1, td2) $\vee$ equivalent (td1, td2)] $\Rightarrow$ includes (x, y)

p(x) $\wedge$ p(y) $\wedge$ name (x, np1) $\wedge$ name (y, np2) $\wedge$ range (x, rp1) $\wedge$ range (y, rp2) $\wedge$ temporal_descriptors (x, td1) $\wedge$ temporal_descriptors (y, td2) $\wedge$ [included in (np1, np2) $\vee$ equivalent (np1, np2)] $\wedge$ [included in (rp1, rp2) $\vee$ equivalent (rp1, rp2)] $\wedge$ [included in (td1, td2) $\vee$ equivalent (td1, td2)] $\Rightarrow$ included in (x, y)

p(x) $\wedge$ p(y) $\wedge$ name (x, np1) $\wedge$ name (y, np2) $\wedge$ range (x, rp1) $\wedge$ range (y, rp2) $\wedge$ temporal_descriptors (x, td1) $\wedge$ temporal_descriptors (y, td2) $\wedge$ [disjoint (np1, np2) $\vee$ disjoint (rp1, rp2) $\vee$ disjoint (td1, td2)] $\Rightarrow$ disjoint (x, y)

Mapping of relations and thematic properties (where f stands for a property or a relation):

f(x) $\wedge$ f(y) $\wedge$ name (x, nf1) $\wedge$ name (y, nf2) $\wedge$ range (x, rf1) $\wedge$ range (y, rf2) $\wedge$ equivalent (nf1, nf2) $\wedge$ equivalent (rf1, rf2) $\Rightarrow$ equivalent (x, y)

f(x) $\wedge$ f(y) $\wedge$ name (x, nf1) $\wedge$ name (y, nf2) $\wedge$ range (x, rf1) $\wedge$ range (y, rf2) $\wedge$ [$\neg$disjoint (nf1, nf2) $\vee$ overlap (rf1, rf2)] $\Rightarrow$ weak_overlap (x, y)

f(x) $\wedge$ f(y) $\wedge$ name (x, nf1) $\wedge$ name (y, nf2) $\wedge$ range (x, rf1) $\wedge$ range (y, rf2) $\wedge$ [$\neg$disjoint (nf1, nf2) $\wedge$ overlap (rf1, rf2)] $\Rightarrow$ strong_overlap (x, y)

f(x) $\wedge$ f(y) $\wedge$ name (x, nf1) $\wedge$ name (y, nf2) $\wedge$ range (x, rf1) $\wedge$ range (y, rf2) $\wedge$ [includes (nf1, nf2) $\vee$ equivalent (nf1, nf2)] $\wedge$ [includes (rf1, rf2) $\vee$ equivalent (rf1, rf2)] $\Rightarrow$ includes (x, y)

f(x) $\wedge$ f(y) $\wedge$ name (x, nf1) $\wedge$ name (y, nf2) $\wedge$ range (x, rf1) $\wedge$ range (y, rf2) $\wedge$ [included in (nf1, nf2) $\vee$ equivalent (nf1, nf2)] $\wedge$ [included in (rf1, rf2) $\vee$ equivalent (rf1, rf2)] $\Rightarrow$ included in (x, y)

f(x) $\wedge$ f(y) $\wedge$ name (x, nf1) $\wedge$ name (y, nf2) $\wedge$ range (x, rf1) $\wedge$ range (y, rf2) $\wedge$ [disjoint (nf1, nf2) $\vee$ disjoint (rf1, rf2)] $\Rightarrow$ disjoint (x, y)

Mapping of mixed properties or relations (where f stands for a property or a relation, and variables with same index are of the same nature, spatial, temporal or thematic):

f(x_1) $\wedge$ f(x_2) $\wedge$f(y_1) $\wedge$f(y_2) $\wedge$ equivalent (x_1, y_1) $\wedge$ equivalent (x_2, y_2) $\Rightarrow$ equivalent ($x_1 \vee x_2$, $y_1 \vee y_2$)

f(x_1) $\wedge$ f(x_2) $\wedge$f(y_1) $\wedge$f(y_2) $\wedge$ [overlap (x_1, y_1) $\vee$ overlap (x_2, y_2)] $\Rightarrow$ weak_overlap ($x_1 \vee x_2$, $y_1 \vee y_2$)

f(x_1) $\wedge$ f(x_2) $\wedge$f(y_1) $\wedge$f(y_2) $\wedge$ [overlap (x_1, y_1) $\wedge$ overlap (x_2, y_2)] $\Rightarrow$ strong_overlap ($x_1 \vee x_2$, $y_1 \vee y_2$)

f(x_1) $\wedge$ f(x_2) $\wedge$f(y_1) $\wedge$f(y_2) $\wedge$ [includes (x_1, y_1) $\vee$ equivalent (x_1, y_1)] $\wedge$ [includes (x_2, y_2) $\vee$ equivalent (x_2, y_2)] $\Rightarrow$ includes ($x_1 \vee x_2$, $y_1 \vee y_2$)

f(x_1) $\wedge$ f(x_2) $\wedge$f(y_1) $\wedge$f(y_2) $\wedge$ [included in (x_1, y_1) $\vee$ equivalent (x_1, y_1)] $\wedge$ [included in (x_2, y_2) $\vee$ equivalent (x_2, y_2)] $\Rightarrow$ included in ($x_1 \vee x_2$, $y_1 \vee y_2$)

f(x_1) $\wedge$ f(x_2) $\wedge$f(y_1) $\wedge$f(y_2) $\wedge$ [disjoint (x_1, y_1) $\vee$ disjoint (x_2, y_2)] $\Rightarrow$ disjoint ($x_1 \vee x_2$, $y_1 \vee y_2$)

Mapping of Views (where f stands for any feature, and V_i for a view) :

$\forall$ f, f$\in$V1, $\exists$ f', f'$\in$ V2| equivalent(f, f') $\Rightarrow$ equivalent (V1, V2)

$\forall$ f, f$\in$ V1, $\exists$ f', f'$\in$ V2| [equivalent(f, f') $\vee$ includes (f, f')]$\Rightarrow$ includes (V1, V2)

$\forall$ f', f'$\in$ V2, $\exists$ f, f$\in$ V1| [equivalent(f, f') $\vee$ includes (f, f')]$\Rightarrow$ included in (V1, V2)

$\exists f,\ f\in V1,\ \exists f',\ f'\in V2 \mid \text{strong\_overlap}(f, f') \Rightarrow \text{strong\_overlap}(V1, V2)$ $[\forall f,\ f\in V1,\ \forall f',\ f'\in V2 \mid [\text{disjoint}(f, f') \vee \text{weak\_overlap}(f, f')]] \wedge [\exists f,\ f\in V1,\ \exists f',\ f'\in V2 \mid \text{weak\_overlap}(f, f')] \Rightarrow \text{weak\_overlap}(V1, V2)$ $\forall f,\ f\in V1,\ \forall f',\ f'\in V2 \mid \text{disjoint}(f, f') \Rightarrow \text{disjoint}(V1, V2)$
Mappings of MVAC (where V_i stands for a view and $MVAC_i$ for a MVAC concept) $\forall V,\ V\in MVAC1,\ \exists V',\ V'\in MVAC2 \mid \text{equivalent}(V, V') \Rightarrow \text{equivalent}(MVAC1, MVAC2)$ $\forall V,\ V\in MVAC1,\ \exists V',\ V'\in MVAC2 \mid [\text{equivalent}(V, V') \vee \text{includes}(V, V')] \Rightarrow \text{includes}(MVAC1, MVAC2)$ $\forall V',\ V'\in MVAC2,\ \exists V,\ V\in MVAC1 \mid [\text{equivalent}(V, V') \vee \text{includes}(V, V')] \Rightarrow \text{included in}(MVAC1, MVAC2)$ $\exists V,\ V\in MVAC1,\ \exists V',\ V'\in MVAC2 \mid \text{strong\_overlap}(V, V') \Rightarrow \text{strong\_overlap}(MVAC1, MVAC2)$ $[\forall V,\ V\in MVAC1,\ \forall V',\ V'\in MVAC2 \mid [\text{disjoint}(V, V') \vee \text{weak\_overlap}(V, V')]] \wedge [\exists V,\ V\in MVAC1,\ \exists V',\ V'\in MVAC2 \mid \text{weak\_overlap}(V, V')] \Rightarrow \text{weak\_overlap}(MVAC1, MVAC2)$ $\forall V,\ V\in MVAC1,\ \forall V',\ V'\in MVAC2 \mid \text{disjoint}(V, V') \Rightarrow \text{disjoint}(MVAC1, MVAC2)$

The *G-MAP Mapping Inference Engine* is illustrated in Figure 3.15. First, MVAC concepts and mappings between Basic MVAC Elements are translated into *MVAC element type statements* (statements that indicate a feature's type, such as p(x)) and *MVAC element mapping statements* that can be compared against the antecedents of mapping rules. These statements are stored in the Fact Base. The Mapping Inference Engine matches facts of the Fact base against rules in the Mapping rule base. If a rule is verified, the inference engine issues the relation stated in the consequent of the rule and stores the produced semantic mapping into the Complex MVAC Element Mapping Repository and in the Fact Base as a new statement. The inference engine verifies another rule until no rules remains in the Mapping Rule Base.

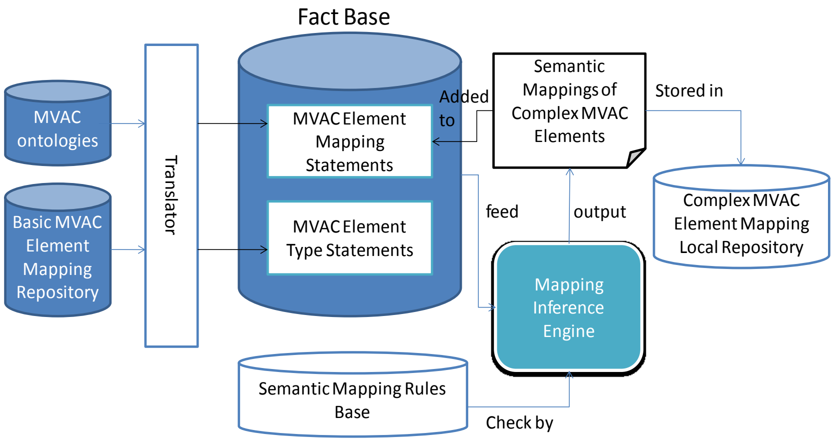

Fact Base
MVAC ontologies
Basic MVAC Element Mapping Repository
Translator
MVAC Element Mapping Statements
MVAC Element Type Statements
Added to
Semantic Mappings of Complex MVAC Elements
Stored in
Complex MVAC Element Mapping Local Repository
feed
output
Mapping Inference Engine
Semantic Mapping Rules Base
Check by

Figure 3.15 *G-MAP Mapping Inference Engine*'s functional architecture

Note that the mappings between spatial and temporal properties depend on the mappings between their spatial and temporal descriptors. Therefore, to optimize the mapping Inference Engine, descriptors are mapped prior to properties.

3.6.3.3 G-MAP Augmented Mapping Inference Engine

The new principle of the augmented mapping inference engine is to exploit the dependencies to discover missing mappings between features of MVACs. For example, consider two properties *depth* of watercourse and *water level*. It is possible that no external resource, such as a lexicon, can discover that they represent the same attribute. However, if we discover that they participate in similar dependencies, we could infer that they may represent the same thing. The reasoning process of the augmented mapping inference engine is illustrated in Figure 3.16. First, the dependencies of MVAC are extracted and stored in Dependency Temporary Repository. In parallel, the system extracts from the Complex MVAC element mapping Repository the non-equivalent pairs of MVAC elements. An assumption is made that the semantic mapping between these elements can be false because implicit information (contained in dependencies) that was not taken into account can influence the result. This assumption is called the "equivalence assumption". For each pair of non-equivalent MVAC elements, the dependencies that contain the elements that are linked by the mapping are selected, and dependencies of the different MVACs are matched, considering the equivalence assumption.

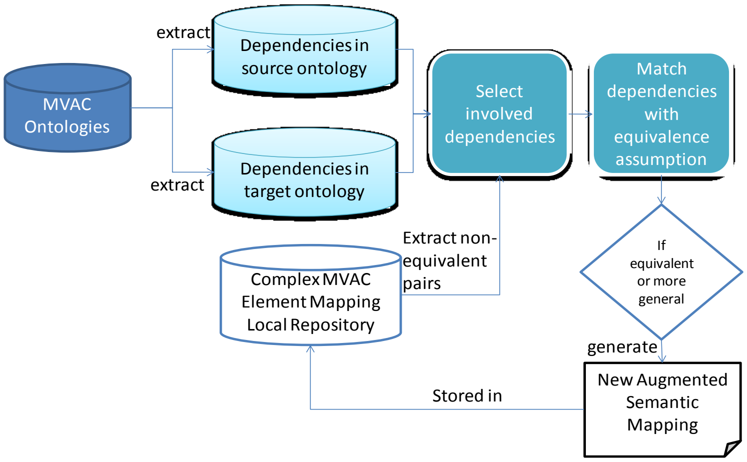

MVAC Ontologies
extract
extract
Dependencies in source ontology
Dependencies in target ontology
Select involved dependencies
Match dependencies with equivalence assumption
Extract non-equivalent pairs
Complex MVAC Element Mapping Local Repository
If equivalent or more general
generate
Stored in
New Augmented Semantic Mapping

Figure 3.16 Augmented mapping inference engine's process

If dependencies are found to be equivalent or one is more general than the other, than the assumption is proven, that is, the pair of elements that was stated to be disjoint is in fact equivalent. The new augmented mapping replaces the previous mapping. For example, consider the following dependencies *d1* and *d2* that belong to different MVAC, *d1*:(*depth*(*stream*, *low*)→ *geometry*(*stream*, *polygon*) and *d2*:*water level*(*watercourse*, *low*)→ *geometry*(*watercourse*, *polygon*) with the following semantic mappings: *equivalent*(*stream*, *watercourse*), *disjoint*(*depth*, *water level*). If we make the assumption: *equivalent* (*depth*, *water level*), we find that *d1* and *d2* are equivalent, and conclude that *equivalent* (*depth*, *water level*) was an implicit mapping. While dependencies represent rich sources of knowledge that can help to discover implicit mappings, results need to be verified by the user since the augmented mapping is based on an assumption.

3.6.3.4 Final Representation of Semantic Mappings between MVAC

The results produced by G-MAP are *MVAC semantic mappings* (Figure 3.17). These mappings are more complex than existing semantic mappings, which usually consist of a single semantic relation, and/or a semantic similarity. The MVAC semantic mapping can be decomposed into a multi-view mapping, where a semantic relation is provided between each pair of views of two different MVACs. The purpose of multi-view semantic mapping is to allow selecting mappings that are valid in a certain context. In addition, the semantic relation can be decomposed into a thematic, spatial and temporal component. For example, the thematic dimension of the semantic mapping is obtained by considering only the thematic features (properties and relations) of the compared concepts during the semantic mapping process. This decomposition in several dimensions is useful to understand the nature of the relation between concepts. Furthermore, in a given context, the user might find relevant only one component of the mapping. For example, if he or she searches for buildings with the purpose of localising their geometry, the thematic component of the semantic mapping is less important than the spatial component; however, if he/she searches for buildings with the purpose of identifying their usage (ex: residential, commercial, etc.), then the spatial component may be less relevant. Note that the semantic similarity between concepts is provided by Sim-Net, which is presented in Chapter 7.

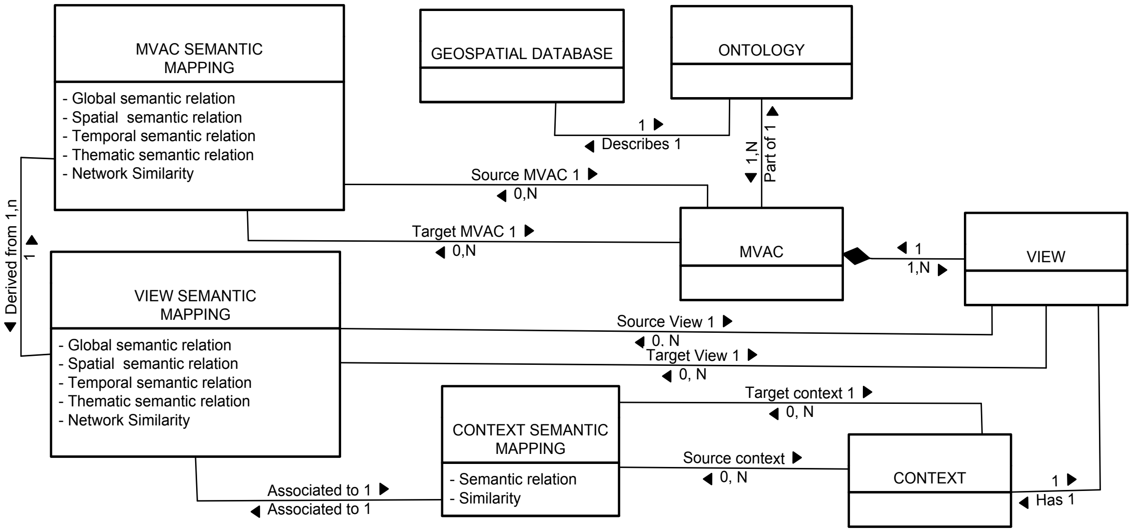


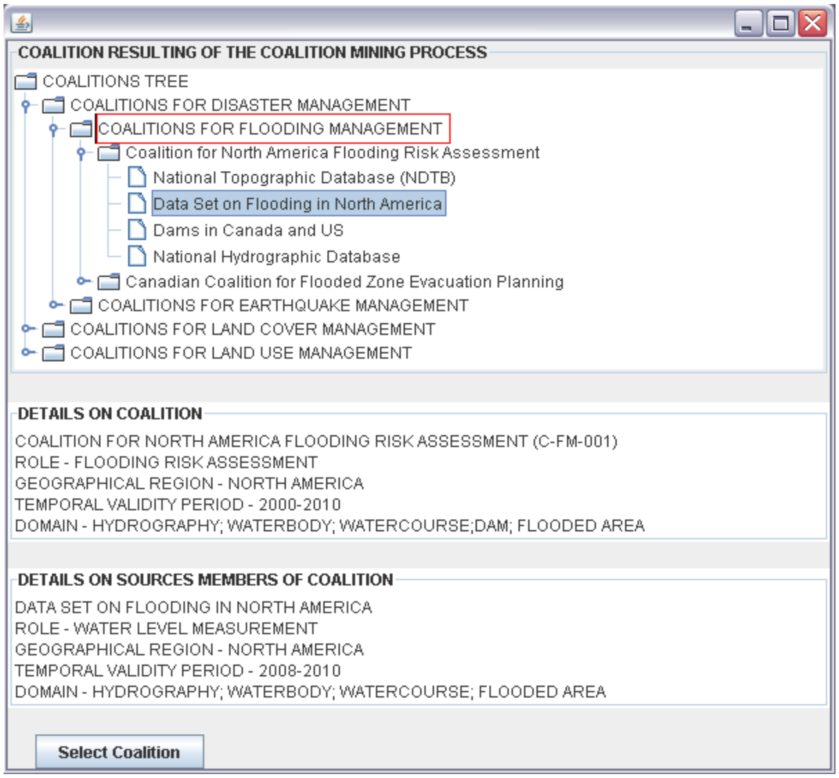
Figure 3.17 MVAC Semantic Mapping Model

Currently, most semantic mapping systems produce semantic similarity values, and very few produce qualitative relations. We argue that even if qualitative relations are often more costly to compute, they provide a better support for the interpretation of shared data.

3.7 Implementation and Experimentation

A Java prototype was implemented based on the proposed framework. The objective of the prototype is to search and discover geospatial databases of the network that can form relevant coalitions, improve the semantic representation of data with the MVAC model, and demonstrate that the G-MAP finds the concepts that are relevant to a given geospatial query and improve the semantic interoperability. The architecture of the prototype comprises three main components: the Coalition Discovery and Visualization component, the MVAC Generation and Visualization component, and the G-MAP Semantic Mapping component. The data used to demonstrate the approach come from various geospatial data sources, including the National Topographic Database of Canada (NTDB), the Quebec

Topographic Database (BDTQ), diverse data sets on disasters (flooding, earthquakes and tornados) in North America, and the Topographic and Administrative Database of Quebec (BDTA). To implement the prototype, we have developed a set of geospatial database's context descriptions, ontologies for databases, and instances, using the OWL ontology language, and based on UML schemas and specifications of samples of the above data sets. The spatial and temporal descriptors of concepts were defined manually based on textual definitions of concepts. We demonstrate the prototype with a scenario where a user searches for *watercourses that are near a given residential area*, in order to *assess flooding risk in diverse regions of North America in the last decade*. First, the user runs the coalition-discovering component that displays the resulting coalitions in a Coalition Tree. The Coalition Tree allows the user to browse the list of coalitions according to a taxonomic classification based in the chosen features: role, domain, geographical location and temporal validity period. The coalition-discovering component integrates a coalition-mining algorithm that is presented in Chapter 4, and it is based on network analysis techniques. In this example, the coalitions are classified according to their function (e.g., flooding management, land use management, etc.) (Figure 3.18).



COALITION RESULTING OF THE COALITION MINING PROCESS

COALITIONS TREE
COALITIONS FOR DISASTER MANAGEMENT
COALITIONS FOR FLOODING MANAGEMENT
Coalition for North America Flooding Risk Assessment
National Topographic Database (NDTB)
Data Set on Flooding in North America
Dams in Canada and US
National Hydrographic Database
Canadian Coalition for Flooded Zone Evacuation Planning
COALITIONS FOR EARTHQUAKE MANAGEMENT
COALITIONS FOR LAND COVER MANAGEMENT
COALITIONS FOR LAND USE MANAGEMENT

DETAILS ON COALITION

COALITION FOR NORTH AMERICA FLOODING RISK ASSESSMENT (C-FM-001)
ROLE - FLOODING RISK ASSESSMENT
GEOGRAPHICAL REGION - NORTH AMERICA
TEMPORAL VALIDITY PERIOD - 2000-2010
DOMAIN - HYDROGRAPHY; WATERBODY; WATERCOURSE;DAM; FLOODED AREA

DETAILS ON SOURCES MEMBERS OF COALITION

DATA SET ON FLOODING IN NORTH AMERICA
ROLE - WATER LEVEL MEASUREMENT
GEOGRAPHICAL REGION - NORTH AMERICA
TEMPORAL VALIDITY PERIOD - 2008-2010
DOMAIN - HYDROGRAPHY; WATERBODY; WATERCOURSE; FLOODED AREA

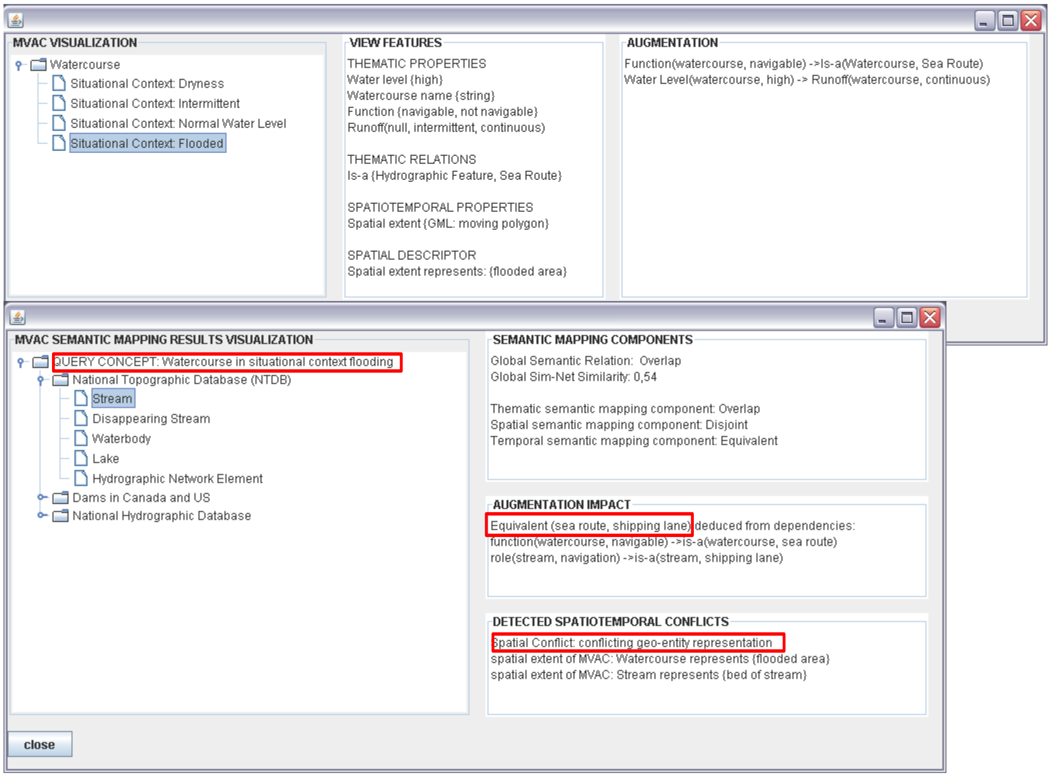
Figure 3.18 Results of Coalitions Tool

The advantage of different classifications is to allow the user to browse the Coalition Tree according to diverse search criteria. To build the hierarchy of roles, we classify all roles of the coalitions according to is-a (subsumption) relation between roles, as derived from external lexical resources. For instance, the category "*coalitions for disaster management*" includes the category "*coalitions for flooding management*".

When a coalition is selected, details on the context of the coalition are displayed, and similarly, details on contexts of coalition members are are also displayed. For example, the coalition for flooding risk assessment in North America is described in OWL RDF/XML syntax as:

```
<Coalition rdf:ID="NorthAmericaFloodingRiskAssessmentCoalition"/>
    <GeographicalLocation rdf:datatype="&xsd;string">NorthAmerica
        </GeographicalLocation>
        <TemporalValidityPeriodStart rdf:datatype="&xsd;gYear">2000
        </TemporalValidityPeriodStart>
        <TemporalValidityPeriodEnd rdf:datatype="&xsd;gYear">2010
        </TemporalValidityPeriodEnd>
        <Role rdf:datatype="&xsd;string">FloodingRiskAssessment</Role>
        <Domain rdf:datatype="&xsd;string">Hydrography</Domain>
        <Domain rdf:datatype="&xsd;string">Waterbody</Domain>
        <Domain rdf:datatype="&xsd;string">Watercourse</Domain>
        <Domain rdf:datatype="&xsd;string">Dam</Domain>
        <Domain rdf:datatype="&xsd;string">FloodedArea</Domain>
</Coalition>
```

The coalitions serve as a basis for supporting collaboration, but also for facilitating the search of relevant geospatial databases. When the user chooses a coalition, before he or she can submit a query, the MVAC tool generates MVACs for the databases' ontologies of the selected coalition, using the method for generation of MVACs described in Chapter 5. For the generation of MVACs, we have manually created the context rules from which contexts of the concept are inferred, based on the databases' specifications. The MVAC tool allows visualizing the MVAC representation of the query concept (ex: *watercourse*) and select the relevant view according to context, for example, "*flooded*", which is a situational context (upper part of Figure 3.19). It also displays the dependencies that augment the concept, for example *function*(*watercourse*, *navigable*)→ *is-a*(*watercourse*, *sea route*) and *water level*(*watercourse*, *high*)→ *runoff*(*watercourse*, *continuous*).



MVAC VISUALIZATION

- Watercourse
  - Situational Context: Dryness
  - Situational Context: Intermittent
  - Situational Context: Normal Water Level
  - Situational Context: Flooded

VIEW FEATURES

THEMATIC PROPERTIES
Water level {high}
Watercourse name {string}
Function {navigable, not navigable}
Runoff(null, intermittent, continuous)

THEMATIC RELATIONS
Is-a {Hydrographic Feature, Sea Route}

SPATIOTEMPORAL PROPERTIES
Spatial extent {GML: moving polygon}

SPATIAL DESCRIPTOR
Spatial extent represents: {flooded area}

AUGMENTATION

Function(watercourse, navigable) ->Is-a(Watercourse, Sea Route)
Water Level(watercourse, high) -> Runoff(watercourse, continuous)

MVAC SEMANTIC MAPPING RESULTS VISUALIZATION

- QUERY CONCEPT: Watercourse in situational context flooding
  - National Topographic Database (NTDB)
    - Stream
    - Disappearing Stream
    - Waterbody
    - Lake
    - Hydrographic Network Element
  - Dams in Canada and US
  - National Hydrographic Database

SEMANTIC MAPPING COMPONENTS

Global Semantic Relation: Overlap
Global Sim-Net Similarity: 0,54

Thematic semantic mapping component: Overlap
Spatial semantic mapping component: Disjoint
Temporal semantic mapping component: Equivalent

AUGMENTATION IMPACT

Equivalent (sea route, shipping lane) deduced from dependencies:
function(watercourse, navigable) ->is-a(watercourse, sea route)
role(stream, navigation) ->is-a(stream, shipping lane)

DETECTED SPATIOTEMPORAL CONFLICTS

Spatial Conflict: conflicting geo-entity representation
spatial extent of MVAC: Watercourse represents {flooded area}
spatial extent of MVAC: Stream represents {bed of stream}

Figure 3.19 MVAC Semantic Mappings Visualisation Tool

The G-MAP produces semantic mappings according to the view selected by the user, making this matching tool more flexible than existing matching tools. The G-MAP tool shows the concepts that were matched with the query concept as a tree, where nodes are databases, and sub-nodes are the retrieved concepts (bottom left of Figure 3.19). When a user selects a retrieved concept, the G-MAP tool displays the thematic, spatial, and temporal components of the semantic mappings, as well as the semantic heterogeneities affecting the spatial and temporal properties that were detected during the mapping process. For instance, while the concept *stream* is matching the concept of *watercourse*, *stream* may not be a relevant concept for the user, considering that the spatial extent of *stream* represents *bed of stream*, while the spatial extent of the query concept *watercourse* represents *flooded area*. The geometries of watercourse and stream objects stored in respective geospatial databases are therefore not comparable. This is a semantic heterogeneity that existing semantic mapping tools do not detect but that is fundamental for the correct interpretation of shared geospatial data. Additionally, the G-MAP tool displays the elements of MVACs that were matched with the augmented mapping functionality (augmentation impact), for example, the match between *sea route* and *shipping lane*. We evaluate the G-MAP semantic mapping algorithm and demonstrate its ability to detect implicit semantic mappings. We used a small network of nine geospatial databases that were developed as described above. Our goal was to evaluate whether the G-MAP would be able to retrieve expected semantic mappings according to a sample of manually identified matching concepts. We employ the F-measure, defined as F-measure = 2*TP/(FN+2*TP+FP), where TP is the number of true positive mappings, FN is the number of false negative mappings, and FP is the number of false positives (Do et al. 2009). The F-measure assesses the ability of the algorithm to retrieve true mappings and reject false mappings. Its value is between 0 and 1, with maximum value indicating a perfect performance. The second measure we used is the overall measure, which evaluates the cost associated with removing false positives and adding false negatives (Do et al. 2009): Overall = 1-(FN+FP)/(FN+TP). This measure gives a value situated between -1 and 1, with -1 indicating a maximal cost and 1 indicating a perfect performance. We tested the algorithm with the complete version of G-MAP (right of Figure 3.20), and without the augmented mapping functionality in order to assess the impact of augmentation (left of Figure 3.20).

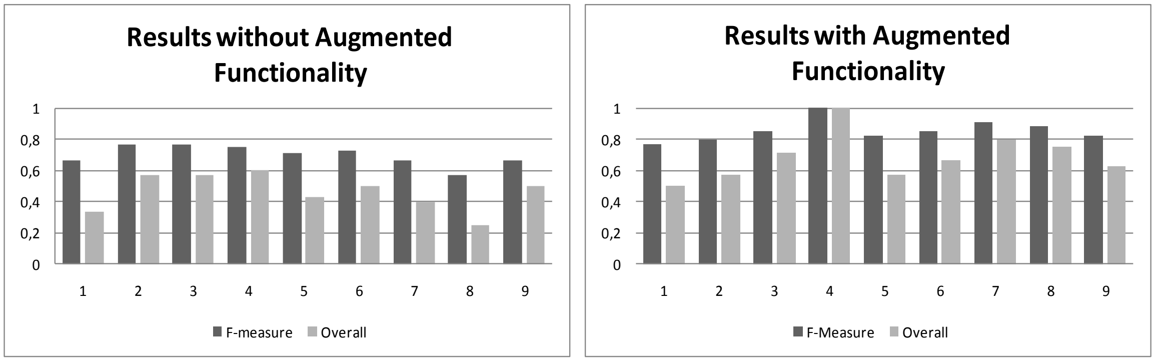


Figure 3.20 Comparative results of G-MAP

The results show that the G-MAP performance, according to the F-measure, is improved when the augmented functionality is employed. However, it is important to note that the performance of G-MAP depends on the richness of the dependencies that augment concepts. A poor set of dependencies will not be enough to increase G-MAP performance. In addition, we note that the representation of dependencies is often lacking in existing geospatial databases.

3.8 Conclusion and Future Work

Semantic interoperability is a problem that has been widely addressed. In ad hoc networks of geospatial databases, many additional issues are raised. The development of dynamic and adaptive approaches is crucial for application domains where the needed data comes from multiple sources, and sources that will have to collaborate are not known in advance. The vision we have given in this paper is that real time semantic interoperability can be based on principles of communication and interactions in social networks. The resulting conceptual framework formalizes the fundamental elements for the development of such an approach. Especially, we have focused on giving a model for geospatial databases coalitions; a new representation of geospatial concepts that includes semantics of spatiotemporal properties, context and dependencies; developing a semantic mapping model associated with this representation and that considers thematic, spatial and temporal components of a concept, which is essential to understand the meaning of shared spatiotemporal data. The implementation of the approach has demonstrated its ability to convey semantics and improve the discovery of matching concepts. The present research opens other research issues. A real time semantic mapping approach is not only a question of automating the discovery of semantic relations among concepts, but also to execute the semantic mappings following a propagation strategy that is determined “on-the-fly”. Future work related to this research will concern the development of the real-time strategies that were highlighted in this framework. Such strategies will be reactive to events that modify the ad hoc network.

CHAPTER 4

Managing Coalitions in Ad Hoc Networks of Geospatial Databases

M. BAKILLAH, M.A. MOSTAFAVI

Submitted to *Data and Knowledge Engineering* journal

4.1 Présentation de l'article

Ayant présenté un cadre conceptuel pour l'interopérabilité sémantique en temps réel dans les réseaux ad hoc de bases de données géospatiales dans le chapitre précédent, les chapitres suivants sont chacun dédiés à un des composants de ce cadre. Dans ce chapitre, nous présentons une approche pour la découverte des sources pertinentes, la formation de coalitions à partir de ces sources pertinentes et la gestion des coalitions dans les réseaux ad hoc. Selon le cadre conceptuel, la découverte et la formation des coalitions constituent la première étape du processus d'interopérabilité sémantique dans un réseau ad hoc de bases de données géospatiales.

Les coalitions de bases de données géospatiales jouent un rôle fondamental pour réaliser l'interopérabilité sémantique, car elles permettent de rassembler les bases de données et leurs utilisateurs qui sont le plus à même d'échanger des données géospatiales et de collaborer. Dans cet article, nous présentons une nouvelle approche pour la découverte et la gestion des coalitions de bases de données géospatiales dans un réseau ad hoc. Nous proposons un modèle décrivant la coalition, lequel inclut une description du contexte et des contraintes thématiques, spatiales et temporelles devant êtres satisfaites pour faire partie de la coalition. Nous proposons ensuite une méthode de découverte des sources pertinentes basée sur un concept d'attraction sémantique entre les bases de données du réseau ad hoc. Nous avons finalement développé un algorithme de fouille de coalitions basé sur des techniques d'analyse de réseau qui permet d'améliorer l'échange de données géospatiales dans le réseau ad hoc dans le contexte d'un processus d'interopérabilité sémantique en temps réel.

4.2 Presentation of the Article

Having presented a framework for real time semantic interoperability in ad hoc networks in the previous chapter, the next chapters each develop a component of this framework. In this chapter, we present the approach for discovering relevant data sources, as well as forming and managing coalitions of geospatial databases in ad hoc networks; this is the first step of the semantic interoperability process according to our framework.

Coalitions of geospatial databases in ad hoc networks play a fundamental role for achieving semantic interoperability, since they allow bringing together those databases and their users that are the most likely to make useful geospatial data exchanges and collaboration. In this paper, we propose a novel approach for the discovery of relevant data sources, the formation and the management of coalitions of geospatial databases in ad hoc networks. A coalition model is proposed which includes a spatial, temporal and thematic description of the coalition context and constraints for being a member of a coalition. We propose a method for discovering relevant geospatial databases using a concept of semantic attraction. We propose a coalition mining algorithm based on network analysis techniques that allows discovering potential coalitions for improving geospatial data sharing in ad hoc networks in the context of a real time semantic interoperability process.

4.3 Introduction

The advent of ad hoc networks has modified the way in which different organizations can share geospatial data and services, allowing for dynamic exchanges and formation of coalitions often for short-term needs. Collaboration is necessary since different organizations, which produce or consume geospatial data, aim at reducing uncertainty and risk in the decision making process by linking with others and taking advantage of shared knowledge (Fox 2008). The concept of collaboration is useful in various areas, including ontology development (Zhdanova 2008) and creating cooperative workflows (Baïna et al. 2004). For instance, in disaster management, different public and private organizations are brought to collaborate together to establish prevention or recovery plans despite their different representations of the same reality. Alhashmi et al. (2008) identifies several collaboration benefits: better results through reduced cycle times, increased sharing of knowledge and technology, and more efficient use of human resources, to name only a few. From the point of view of semantic interoperability, and especially in ad hoc networks, the role of coalitions is to bring together those databases and their users that are the most likely to make useful geospatial data exchanges. More specifically, the role of coalitions is to dynamically aggregate databases with similar knowledge or complementary interests; this will facilitate the discovery of relevant databases and support the propagation of queries issued by nodes to the relevant databases. For the development of the Semantic Web, the interrelation between concepts of collaboration and interoperability is crucial (D'Aquin et al. 2008).

The research question is how can we discover databases of a network that can form coalitions for achieving a common goal? Approaches for supporting and managing coalitions, or collaborations, from a semantic-based point of view are at their beginning, and few approaches on this topic have been proposed in the literature. Until now, nodes of a network have been organized and aggregated into groups in a manual fashion, i.e. by administrators and application experts (Kantere et al. 2008). Because this approach requires manual work, it is not scalable and efficient for large networks. Therefore, coalitions should be discovered and managed dynamically with automatic techniques. Among the approaches that have been proposed regarding this issue, most of existing approaches are based on exact string-matching of the interests of the nodes that may participate in a coalition; more particularly, they do not take into account the spatial and temporal aspects of coalitions (Agostini and Moro 2004; Bloehdorn et al. 2005; Khambatti et al. 2002; Mika 2004). In this context, the contribution of this paper is to propose a semantic-based coalition discovery and management approach that is adapted to geospatial databases. The database discovery process is based on a new concept of semantic attraction which finds pairs of databases whose contexts attract each other. The coalition mining algorithm we have developed for discovering coalitions is based on network analysis techniques and allows discovering databases which act as “central attractors” for other databases. In order to manage coalitions, a set of coalition operators is proposed, which describe the various changes, or events, that coalitions can undergo.

The paper is structured as follows: the next section presents related work on the broader topic of group formation; we focus on content-based approaches, which are closer to the one we propose. Section 4.5 presents the basic definitions that support our approach, including the definition of the coalition model. In section 4.6, the concept of semantic attraction between databases is detailed. Section 4.7 presents the coalition mining algorithm, as well as the operators for managing the coalitions. The implementation and experimentation are presented in section 4.8. Finally, a conclusion and an overview of future work are offered in section 4.9.

4.4 Related Work

The problem of group formation can be defined as the problem of gathering members of a network (which can be sources, peers, agents, etc.) into meaningful groups. We consider that there are two main categories of approaches toward group discovery and formation:

- Agent-based group formation approaches: these approaches are targeted at multi-agents systems (MAS). They are based on the interactive capabilities of agents. Groups are considered as the result of a negotiation among agents, which is guided by notions taken from game theory, such as gain (what an agent “earns” by entering a group) and payoff (what an agent “pays” or “looses” when entering a group). Examples of such approaches are (Dang et al. 2003; Sauro 2005; Ovarec et al. 2007; Zheng et al. 2008; Boella et al. 2009; van der Torre et al. 2009). Agent-based group formation approaches do not necessarily consider the problem of discovering agents that can be part of the group, but focus on providing the agents with functionalities that enable them to reproduce a negotiation process.

- Content-based group formation approaches: in these approaches, groups are formed based on a common knowledge or a common interest among members. In this work, we are closer to these approaches, which are mostly concerned with the problem of discovering nodes of the network that can be part of a group.

The idea of group formation emerged in the context of Peer-to-Peer (P2P) systems, with some approaches proposing the concepts of interest groups (Giunchiglia and Zaihrayeu 2002), P2P communities (Khambatti et al. 2002; Crespo and Garcia-Molina 2002; Castano and Montanelli 2005; Liu et al. 2006) and Peer Federations (Bonifacio et al. 2002). Peer groups are useful to structure the search space and therefore support efficient discovery of resources. According to (Khambatti et al. 2002), P2P communities are sets of peers that share a common interest; this common interest corresponds to the intersection of keywords expressing interest of participating peers. Keywords belong to a common vocabulary that each peer agrees with; therefore, the problem of semantic heterogeneity between peers' interests is avoided. However, assuming that all peers of a network can commit to a common vocabulary is not appropriate for large and heterogeneous networks. P2P communities can also be formed of semantically related peers, that is, peers that hold similar knowledge (Crespo and Garcia-Molina 2002; Löser et al. 2003). In (Castano and Montanelli 2005), the topic of interest of a peer, called Community Identity Card (ICard), is defined as an ontology or concepts and properties. Peers submit discovery queries to find the peers that could be part of an interest-based community. A semantic matchmaker, H-MATCH (Castano et al. 2003) is used to verify if two peers have a common interest and resolve semantic heterogeneity. In this approach, only peers that have similar knowledge can be part of an interest-based community. However, we argue that groups can be formed with members that may have dissimilar knowledge but, for example, want to achieve a common goal or have complementary roles. Liu et al. in 2006 have proposed an approach that is similar to that of Castano and Montanelli, however they also introduce a privacy management scheme where cryptographic protocols are employed to measure similarity between peers without disclosing their personal profiles. Profiles are represented as feature vectors, and similarity is a function of scalar product between feature vectors. In (Crespo and Garcia-Molina 2002), semantically related peers are grouped to form Semantic Overlay Networks (SONs), and SONs are organized in a hierarchy. Similarly, in the work of

Lumineau and Doucet (2004), a community is defined by a small set of keywords that define an area of interest, and interests are organized in a local hierarchy of interest (formed for a small subset of the P2P network), in opposition to SONs which are organized in a global hierarchy common to the whole network. With respect to interest-based groups, SONs gather peers that hold data about similar topics; this is useful to support peer discovery, but it is not a suitable approach for supporting collaboration, since collaboration often implies people holding different but complementary data to work together, not necessarily people just holding similar data. In our approach, we argue that there should be several criteria for forming groups. Furthermore, the existing interest-based peer groups are poorly defined with only keywords. We argue that the semantics of such groups can be better defined with a structured context that would comprise several parameters. Furthermore, existing group formation approaches are not adequate for geospatial databases, since the representation of the semantics of a group does not represent spatial and temporal features of groups. Consequently, there is a need to define the notion of groups of geospatial data sources, and an adapted group formation framework.

4.5 Geospatial Database Context and Coalitions

In a network, coalitions of databases are formed to facilitate collaboration among users of these databases. To determine which databases can form coalitions, background information about them is required. In our approach, this background information is modelled as the context of the database. In this section, we start by proposing an ontological context model for geospatial databases. There are different representations of the context in the literature, and each has been developed for a particular aim:

- the words surrounding a group of words in a text (Finkelstein et al. 2001); the purpose of this context representation is to help in the disambiguitation of a word.

- the description of the environment and situation in which a device (such as mobile sensor) is used (Dey 2001); this representation of context is mostly used in context-aware systems.

- a local model expressing the point of view of a community or an individual (Serafini et al. 2003); according to this definition, the context can be represented by the ontology that was developed to describe a domain of interest.

- personal data about the user, such as its interest and profile (Lawrence 2000; Firat et al. 2007).

In our approach, we define the context of databases as follows: the context is the information that can be used to characterize geospatial databases and that is considered to be relevant to determine the interaction between these databases. Because of the wide variety of geospatial databases, formalizing geospatial database context in a comprehensive fashion is a difficult task. Basically, the context should include the spatial, temporal and thematic aspects that can best describe the data in a succint fashion. We consider that the following context parameters are relevant to define the context of a geospatial database: the geographical location, the temporal period, the domain and the function; these parameters will represent the main concepts of the ontological context model. The model is illustrated in Figure 4.1. The geographical location is the geographical area covered by instances (spatial entities) stored in the database. It is a fundamental context parameter, since members of a coalition often collaborate to resolve a problem in a given geographical location (e.g. in case of a disaster, land management in a given city, etc.). This geographic location can be represented by a *place*, which is an "area of the Earth's surface that possesses some form of identity," and that can be included in a *gazetteer* (Goodchild 2009) (p. 18). A gazetteer defines relations between places, their locations, and their types (Goodchild and Hill 2008). More specifically, it can define inclusion relations between places (places that are located inside other places) (for example, see Fu et al. (2005) or the SWETO-GS ontology of places of Arpinar et al. in 2006). The role of this spatial relation is to determine spatial inclusion between geographical locations covered by different databases when contexts of those databases are compared. The temporal period parameter is the historical period of time when instances stored in the database exist; this parameter is fundamental as well, since it can be imposed as a criterion to enter a coalition, and it is more likely that coalitions contain databases describing the same period. We model time as a temporal period, and instant (date) or an event. The event is useful because in some cases it may be more meaningful than a date. For example, a database on "hurricane Katrina" is more meaningful than a database on August 2005.

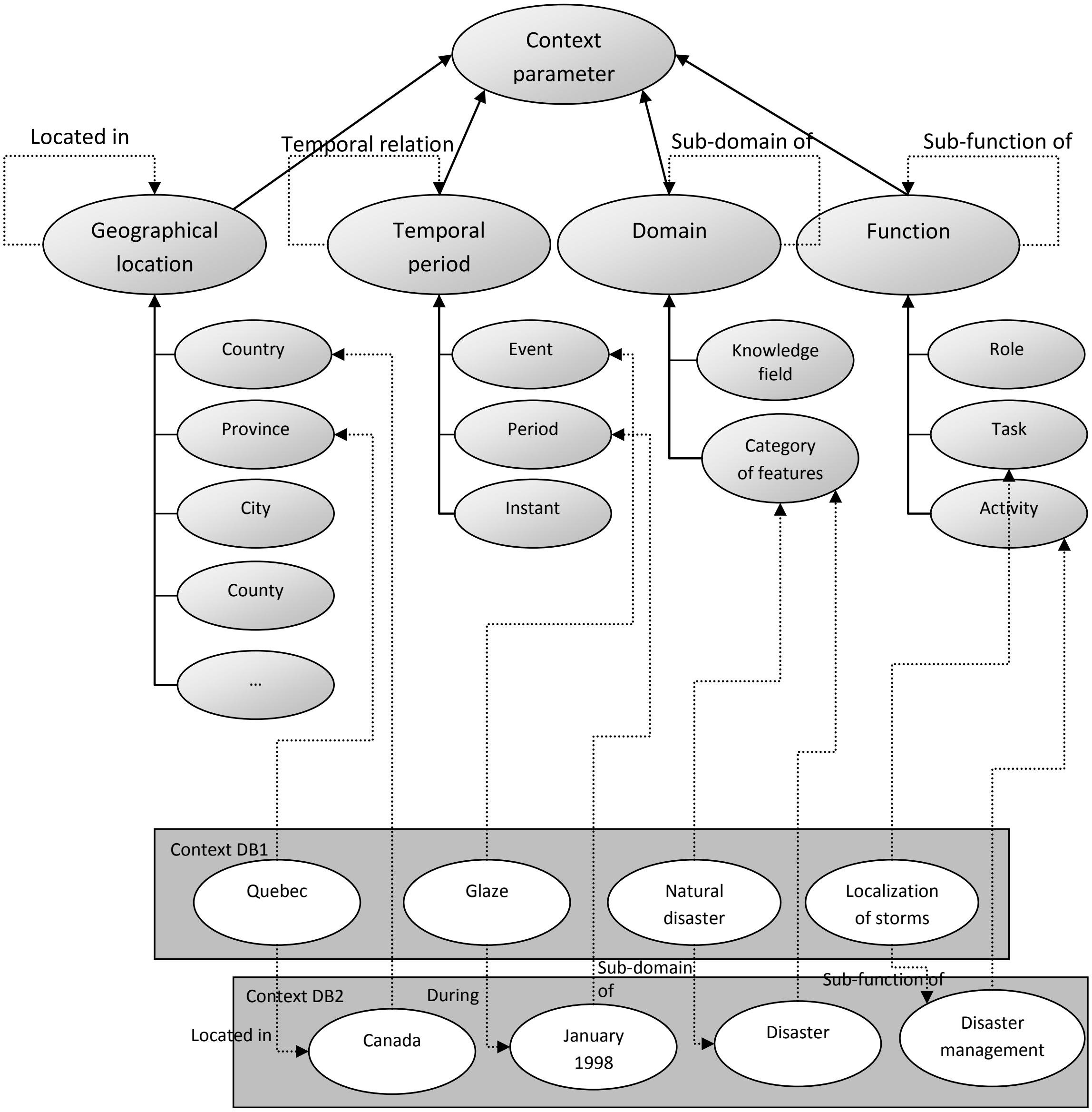


Figure 4.1 Ontological Context Model

In addition, time elements can be related with temporal relations, which include Allen's temporal relations (1983), and additional temporal relations such as those included in Open Cyc Temporal Ontology, as decribed in Section 3.6.3.1.1. The role of temporal relations is similar to that of the inclusion relation between geographical locations. Domain is the area of knowledge that the database is related to; it can correspond to a domain area (hydrology, ecology, urbanism, etc.) of a category of features (rivers, buildings, floods, etc.). The domain of a database should encompass the main categories of geographical entities referred to by the concepts that compose the ontology that describes the semantics of this database. The function is the main reason for which the database was built; it can be a role, a task or an activity (e.g. waste management, land use assessment, contamination risk assessment, etc.). The function of the database corresponds to the task that the user wants to perform with the data. This parameter is fundamental, since, according to Wiegand and Garcia (2007), many searches for geospatial data are based on the intended tasks, for example, during a response to a disaster or for land use or urban planning. Domains and function are related through "sub-domain of" and "sub-function of" relations, which will allow matching different but related domains and functions of different databases.

Databases' contexts are instances of the context model. An example is shown at the bottom of Figure 4.1, where short dashed arrows represent a relation between an instance and the concept it instantiates. Therefore, a database context is defined as a tuple $ctx(DB)$ = <*GL(DB)*, *T(DB)*, *DOM(DB)*, *F(DB)*>, where *GL(DB)* is the geographical location, *T(DB)* is the time, *DOM(DB)* is the domain and *F(DB)* is the function.

In addition to a database and its context, we consider that each node of the network can be associated with a user. Users may be interested in forming or being part of specific coalitions of geospatial databases. Each user can represent its interest by specifying one or several values for each of the context parameter (domain, function, geographical location, and temporal period). Therefore, the user's interest amounts to the description of the coalition that the user of database *DB* would like to be part of. To express the user's interest, we use spatial, temporal and thematic constraints for the context parameters. These constraints are defined below and schematically represented with constraint diagrams (Kent 1997):

Definition 1 (spatial coalition constraint). A spatial coalition constraint *s* is a constraint on the geographical location covered by the coalition: *s* = <G(C) op IG>, where G(C) is the geographical location of the coalition, op is a comparison operator (=, ⊆, ∩, etc.), and IG is a geographical location, with ¬(G(O) ⊥ IG) (where ¬ is the negation operator and ⊥ means "disjoint") .

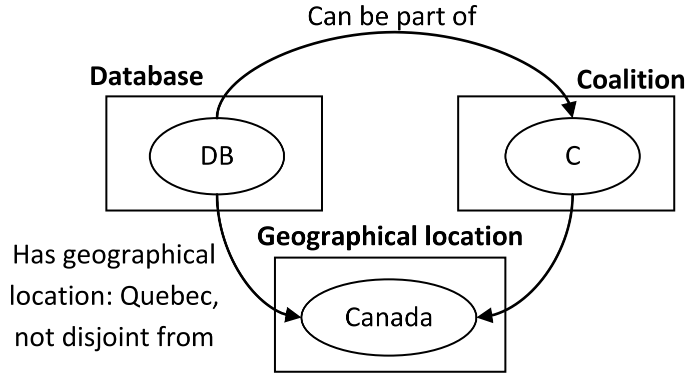


Figure 4.2. Example of spatial coalition constraint

Definition 2 (temporal coalition constraint). A temporal coalition constraint t is a constraint on the temporal validity period of the coalition: t = <P(C) op IP>, where P(C) is the temporal validity period of the coalition, op is a comparison operator (=, ⊆, ∩, etc.), and IP is a temporal validity period, with ¬(P(O) ⊥ IP).

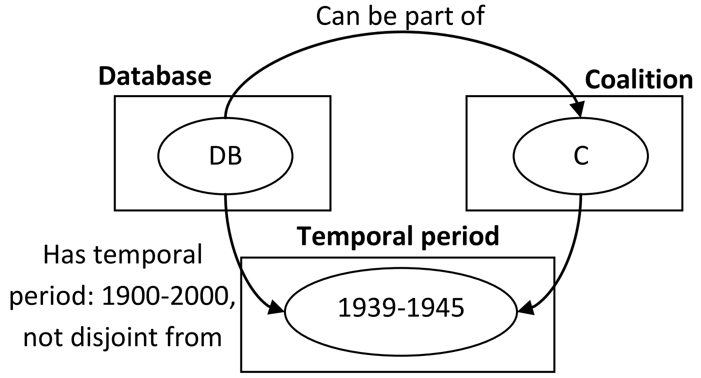


Figure 4.3 Example of temporal coalition constraint

Definition 3 (thematic coalition constraint). A thematic coalition constraint th is a constraint on the domain or on the function of the coalition: th = <D(C) op ID>, where D(C) is the domain of the coalition, op is a comparison operator (=, ⊆, ∩, etc.), and ID is a domain label, with ¬(D(O) ⊥ ID), or th = <F(C) op IF>, where F(C) is the function of the coalition, op is a comparison operator (=, ⊆, ∩, etc.), and IF is a function label, with ¬(F(O) ⊥ IF).

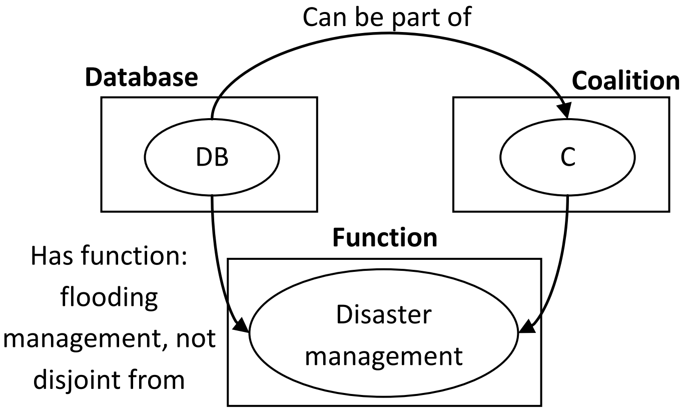


Figure 4.4 Example of thematic coalition constraint

While geographical location, temporal validity time, domain and function are all static variables (they were fixed at the database design time), the interest is a dynamic variable that may change as the user wants to be part of different coalitions.

In our approach, we rely on an upper-level ontology that defines a common vocabulary describing domains and functions and that will allow comparing elements of the contexts of different databases on a common ground. An upper-level ontology is a domain independent ontology that contains general concepts at a high level of granularity (Kavouras 2003). The upper-level ontology of domains and functions organizes knowledge about domains and functions in terms of concepts, taxonomic relations, and semantic annotations:

Definition 4 (Upper-level ontology of domains and functions DFO). DFO is a tuple of the form (*C*, *R*, *A*), where:

- *C = {c1, c2...}* is a set of concepts describing domains or functions;
- *R= {r1, r2...}* is a set of taxonomic relationships (i.e., is-a relations) among concepts of *C*. A taxonomic relationship is a binary relation of the form r(c, c'), where c is more general than c'.
- *A= {a1, a2...}* is a set of semantic annotations relating an element *ε* (*function* or *domain*) of the context of a database *DB* to a concept *c* ∈ *C*. An annotation *a* ∈ *A* writes as a binary relation *a*(*ε, c*).

Semantic annotation is used to link an information source to an element of an ontology (Lemmens 2006). In our context, it allows defining the semantics of this element (*domain* or *function*) according to a common vocabulary for the network. When an annotation is stored in a local ontology, it is also referred to as a registration (Bowers and Ludäscher 2004). We assume that when a database is introduced in the network, a new context description (i.e. an instance) is created using metadata of the database, and relations between context parameters (sub-domain of and sub-function of) are created (relations can be identified with the help of external resources such as terminological databases). The DFO is updated accordingly.

Definition 5 (Coalition). A coalition of databases *Π = <DB(Π)*, *C(Π)*, *CTX(Π)>* is a collection of databases *DB(Π)* that verifies a set of spatial, temporal and thematic coalition constraints *C(Π)*. The spatial, temporal and thematic coalition constraints of the coalition are defined as the intersection of all constraints defined in the interest of coalition members. Once the coalition is established, the coalition constraints can be used to express conditions that must be met by databases to enter a coalition. The context of the coalition *CTX(Π)* is given by *CTX(Π)=<GL(Π)*, *T(Π)*, *DOM(Π)*, *F(Π)>*. *GL(Π)* is the geographical location covered by the coalition, defined as the union of the geographical locations covered by databases in *DB(Π)*. *T(Π)* is the temporal period of the coalition, defined as the union of the temporal periods of databases in *DB(Π)*. *DOM(Π)* and *F(Π)* are respectively the domain and function of the coalition. We determine the domain and function of the coalition by merging the corresponding variables of member databases:

Definition 6 (Domain Merging). The domain of a coalition is given by the union of domains of all participating databases:

$$DOM(\Pi) = \bigcup_i D(DB_i)$$

where $DB_i \in DB(\Pi)$. We define the global domain of a coalition as the union of local domains because we argue that it is the sum of all knowledge of a group that forms its global knowledge.

Definition 7 (Function Merging). The function of a coalition is given by the union of functions of all participating databases:

$$F(\Pi) = \bigcup_i F(DB_i)$$

where $DB_i \in DB(\Pi)$. The semantics of the union operator is defined as follow: if the set of domains (functions) of the databases forming a coalition is the same as the set of the sub-domains (sub-functions) of a more general domain (function) in the DFO, then the merging of the domains of databases correspond to this more general domain. Otherwise, the merging of domains (function) is simply the set of individual domains (functions) of member databases.

4.6 Semantic Attraction for Discovering Geospatial Databases

In this section, we present the notion of semantic attraction between databases, which is used to discover the geospatial databases that can be gathered to form a coalition. Semantic attraction is a measure of how databases are “attracted” to each other based on the similarity between elements of their respective context. Beside the similarity of their contexts, semantic attraction between databases also depends on the interests of their users (as expressed by constraints defined in Section 4.5). Semantic attraction is a function of the semantic distance among databases, which is defined in the following.

4.6.1 Distance Between Geospatial Databases

We define different distances between databases, depending on the elements of their context being compared: distance relative to geographical location (i.e., the geographical area covered by the instances stored in the database), denoted with $\delta_G(DB_i, DB_j)$, distance relative to temporal period, $\delta_T(DB_i, DB_j)$, distance relative to domain, $\delta_{dom}(DB_i, DB_j)$, and distance relative to function, $\delta_F(DB_i, DB_j)$. When measuring the semantic distance between databases, we consider two different cases. In the first case, (1) a database DB_i is attracted by another database DB_j because the elements of context of DB_i match the interest (constraints) of the user of DB_j (the attraction is unilateral); in the second case, (2) a database DB_i is attracted by another database DB_j because they have similar contexts elements (this attraction is multilateral). Based on these two cases, we have elaborated two ways of measuring the semantic distance:

Unilateral coalition scheme (UCS). This scheme is referring to the first case, i.e., when two users of databases DB_i and DB_j respectively accept to be part of the same coalition even though only one of the users (e.g., user of DB_i) has some interest in the coalition. Let $\varepsilon_k(DB_i)$ be an element of the context of database DB_i (with ε_k = geographical location, temporal period, domain or function). Let $I_k(DB_i)$ be a constraint expressing the interest of DB_j's user with respect to one of the context elements (geographical location, temporal period, domain or function). In the unilateral coalition scheme, the semantic distance between two context elements of the same type is defined as the minimal value between:

- the distance between context element $\varepsilon_k(DB_i)$ of DB_i and the corresponding context element $I\varepsilon_k(DB_j)$, and
- the distance between context element $\varepsilon_k(DB_j)$ of DB_j and the value of corresponding context element $I\varepsilon_k(DB_i)$ in coalition constraints of DB_i:

$$\delta_k(DB_i, DB_j) = \min[\delta(\varepsilon_k(DB_i), I\varepsilon_k(DB_j)), \delta(\varepsilon_k(DB_j), I\varepsilon_k(DB_i))]$$

where k = Geometrical location GL | Time period T | Domain DOM | Function F.

Multilateral coalition scheme (MCS). This scheme is referring to the second case where users accept to form a coalition only if all users have some interest in it. In this case, the distance will be simply defined as the distance between value of context elements $I\varepsilon_k(DB_i)$ and $I\varepsilon_k(DB_j)$:

$$\delta_k(DB_i, DB_j) = \delta(I\varepsilon_k(DB_i), I\varepsilon_k(DB_j))$$

The way the distance is measured depends on the elements compared. For function and domains, distance is evaluated by considering how close they are in DFO. The distance δ(c, c') between two concepts (nodes) c and c' in DFO is the length (number of relations) of the shortest path between those nodes in DFO (Rada et al. 1989). The distance between two geographical locations G and G' (defined in different contexts) can be measured in a similar way, by using an ontology of places, or gazetteer, instead of the DFO ontology. The distance between G and G' is the length (number of relations) of the shortest path between those geographical locations in the gazetteer. Finally, to measure the distance δ_t between two time periods T and T', we need to use a different distance, because the compared elements are intervals that are not situated within a hierarchy such as that being defined in the DFO. The distance δ_t between two time periods T and T' is measured with the Minkowski distance (for a more complete definition of this distance, see Schwering (2006), p. 65-66). The general expression of Minkowski distance between two elements *i* and *j* situated in a space with *n* dimensions is given by:

$$d_{ij} = \left[\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^r \right]^{1/r} ,$$

where x_{ik} is the value of dimension *k* for element *i* and x_{jk} is the value of dimension *k* for element *j*. r=1 is the city-block distance and r=2 is the Euclidean distance. We choose r=2 to avoid negative values. Applying this distance to measure the distance between time interval (*n*=2), we have:

$$\delta_t(T(C1),T(C2)) = \left[|t_{11}-t_{12}|^2 + |t_{21}-t_{22}|^2\right]^{1/2},$$

Where t_{11}, t_{12} are the start and end of time period T(DB1) of database DB1, and t_{21}, t_{22} are the start and end of time period T(DB2) of database DB2.

These values of distance are incorporated into the definition of the semantic attraction measure.

4.6.2 Semantic Attraction Between Databases

Consider two databases DB_i and DB_j, with respective context ctx(DB_i) and ctx(DB_j). Consider that we have computed the semantic distance between the four pairs of context elements of these databases, as explained in the previous section; these distances are denoted with δ_k, k = GL | T | DOM | F. Let φ_{ik} and φ_{jk} be the corresponding weights given to the four different types of context parameters, with each weight comprised between 0 and 1; φ_{ik} is the weight given by user of database i to the context element of type k, while similarly, φ_{jk} is the weight given by user of database j to the context element of type k. The idea of the semantic attraction is to integrate the impact of the weights given by the different users with the semantic distance into a formula that is analogous to the gravitational attraction between two physical bodies: the weights provided by the users are analogous to the masses of the two physical bodies, while the semantic distance is analogous to the distance between the bodies in space. Based on this analogy, we define the semantic attraction between databases as follows:

$$F(DB_i, DB_j) = \sum_k \frac{\varphi_{ik} \times \varphi_{jk}}{\delta_k^{\,2}(DB_i, DB_j)}$$

The semantic distance between context elements influences the semantic attraction as follow: the semantic attraction decreases linearly with the square of the semantic distance between databases. That is, the more the semantic distance among databases is high, the more the semantic attraction is low. This formula can be used in two ways. To discover relevant databases, we may use it to compare a representation of a virtual database'context, which is the query, with real contexts of databases in the network. To directly discover coalitions, we can use it in the coalition mining algorithm defined in Section 4.7.

4.6.3 Semantic Attraction Between Coalitions

Consider two coalitions Π_1 and Π_2 formed respectively with sets of databases *DB*(Π_1) and *DB*(Π_2). We define the semantic attraction among coalitions to be:

$$F(\Pi_1, \Pi_2) = \sum_i \sum_{j \neq i} F(DB_i, DB_j)$$

with $DB_i \in DB(\Pi_1)$ and $DB_j \in DB(\Pi_2)$. We can also compute the semantic attraction considering the contexts of the coalitions.

4.7 Managing Coalitions

The semantic attraction measures described in the previous section are used to manage the coalitions of databases which have acquaintances with respect to their contexts. Managing coalitions involves the following tasks:

- Discovering potential groups of databases that can form coalitions. This is done with a **coalition mining algorithm** we have developed;

- Managing various types of **coalition events**, including: (i) forming coalitions; (ii) expanding coalitions with new databases; (iii) coalition shrinking (removing databases from existing coalitions); (iv) merging coalitions in bigger coalitions; (v) dividing a coalition in several smaller coalitions; (vi) dissolving an existing coalition.

4.7.1 Coalition Mining Algorithm based on Network Analysis

We have developed a coalition mining algorithm which uses semantic attraction to identify sets of databases that could form a coalition. The algorithm is based on methods of network analysis, which are methods for discovering structures in various types of networks (Hoser et al. 2006). For instance, network analysis is used to discover groups of communication in social networks (Fox 2008; Pathak et al. 2007). The principle of the proposed algorithm is the following: first, semantic attraction is computed between couples of databases of the network. These semantic attraction values are stored in an adjacency matrix. The adjacency matrix is analyzed to identify "special" databases that attract more databases than other, and seem to play the role of "dominant" databases in the network. We call such database an *attractor database*. For each attractor database, a coalition will be created. To identify the databases that will be part of the coalition of a given attractor database, we iteratively search the semantic neighborhood of the attractor database to find other databases whose semantic attraction toward the attractor database is significant. The coalition mining algorithm also resolves the issue of semantic heterogeneities between databases contexts of users' interest through the semantic attraction measurement, since semantic attraction is able to identify pairs of context elements which are not exactly matching but are semantically related.

The input of the coalition mining algorithm is a set of n databases that form the network (or it could be a subset of it). The output is the set of potential coalitions. First, we randomly compute semantic attraction among couples of databases (step 1 of the algorithm). The computed values are stored in the adjacency matrix:

$$A(t) = \begin{pmatrix} - & F_{12}(t) & \ldots & F_{1n}(t) \\ F_{21}(t) & - & \ldots & \ldots \\ \ldots & \ldots & - & \ldots \\ F_{n1}(t) & \ldots & \ldots & - \end{pmatrix}$$

where F_{ij} is a short notation for $F(DB_{\mathrm{i}}, DB_{\mathrm{j}})$, and t is the time. The adjacency matrix is symmetric because the semantic attraction formula is symmetric. Its size is $n \times n$. By default, semantic attraction between a database and itself is not defined and indicated with "-". After a given number of computations, we stop the computation process to evaluate if we can identify some attractor database in the current state of the adjacency matrix $A(t)$. We set the number of computations to nc. This parameter must be large enough to enable computing a representative sample for the network, but not too large if we want the method to be scalable. To determine if a database is an attractor, we use the degree of centrality index (Hoser et al. 2006) (step 2). This index is an indicator of the density of links that emanates from a node in a network. The higher the degree of centrality of a database, the more we can consider this database as an attractor database. The degree of centrality of database DB_k is given by the following formula:

$$DC(DB_k) = \sum_{l=1}^{n} F_{kl}$$

To evaluate whether the degree of centrality of a database DB_k is high enough to consider it as an attractor database, we evaluate the deviation of $DC(DB_\mathrm{k})$ from the mean value of DC for all databases. The mean value of DC, denoted by μ, is given by the following formula (step 3):

$$\mu = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} F_{ij}$$

The deviation of $DC(DB_\mathrm{k})$ from the mean value is given by:

$$\Delta(DB_k) = (DC(DB_k) - \mu)$$

We evaluate the deviation only for databases whose degree of centrality DC is higher than a given minimum DC_{min}, in order to avoid useless computation for databases with too low degree of centrality (step 4). If a database shows deviation $\Delta(DB_\mathrm{k})$ higher than a given threshold value Δ_min, then we designate it as an attractor database (step 5). If no attractor database is detected at this stage, it means that for the computed sample the semantic attraction were either relatively homogeneous or very low (lower than allowed by the threshold DC_{min}). Based on this sample we could not, therefore, detect an attractor database. We go back to the beginning of the algorithm to compute more semantic attractions and start over the procedure with a larger sample (step 6). If the adjacency matrix is completely filled and no attractor database was detected, we may have to re-evaluate the value of thresholds DC_{min} and Δ_min.

When attractor databases are detected, for each of them we will form a potential coalition (step 7). To select the databases that will be part of the coalition, we follow an iterative process where we examine the successive neighbourhood layers of the attractor database. We call these i^{th} neighbourhood layer the i^{th}-order neighbourhood (Figure. 4.5).

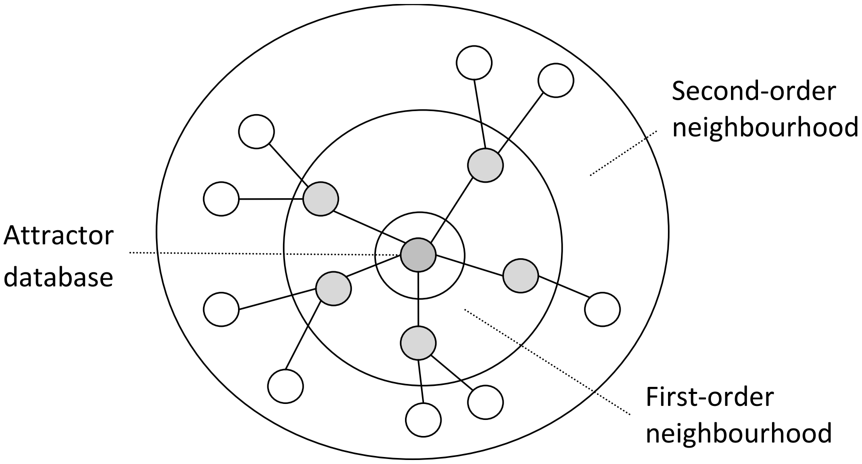


Figure 4.5. Neighbourhoods of an attractor database in the network. Links between nodes (representing databases) represent semantic attractions.

The first order neighbourhood of an attractor database (*FON*(*i*)) is the set of databases for which semantic attraction with DB_i is higher than the semantic attraction threshold F_{th}. The second order neighbourhood of an attractor database (*SON*(*i*)) is the set of databases for which semantic attraction with a database of *FON*(*i*) is higher than the semantic attraction threshold F_{th}. In general, the *k*-order neighbourhood of an attractor database (*k*-*ON*(*i*)) is the set of databases for which the semantic attraction with a database of the previous neighbourhood (i.e., (*k*-1)*ON*(*i*)) is higher than the semantic attraction threshold F_{th}. A database of the i^{th}-order neighbourhood is added to the potential coalition if its semantic attraction with the attractor database is higher than the semantic attraction threshold F_{th}. One of the roles of the *k*-order neighbourhood is to indicate how close from the attractor database a database is. The coalition mining algorithm goes as follow:

Coalition Mining Algorithm

```
Input: a set of n database {$DB_i$| $i$ = 1,…, $n$}
Output: a set of potential coalitions {$\Pi_1$, $\Pi_2$, …}

Initialization:
```

```
- Adjacency matrix $A$, with $A_{ij} = 1$ for i=j
- $nc$ the number of computation at each cycle
- Centrality Vector of centrality $CM$
- Centrality mean $\mu = 0$
- $DC_{min}$ the minimal degree of centrality to be an attractor database
- $\Delta_{min}$ the threshold deviation for degree of centrality
- $A_{th}$ the threshold attraction
- m the maximal number of iteration ($m^{th}$ order-neighbourhood)

Step 1: Randomly compute attraction among sample of databases and
store them in the adjacency matrix A
Step 2: After $nc$ computations, compute the degree of centrality
$DC(DB)$ for each database $DB$
Step 3: Compute the centrality mean $\mu$ for adjacency matrix A
Step 4: Compute the deviation $\Delta(DB)$ for each database $DB$ where $DC(DB)$
$> DC_{min}$
Step 5: If $\Delta(DB) > \Delta_{min}$, select $DB$ as an attractor
Step 6: If no database is selected as an attractor database, return
to step 1
Step 7: For each attractor database $DB_i$, define the coalition $\Pi_i$ with
$DB(\Pi_i) = \{DB_i\}$
Step 8: For each attractor database $DB_i$, define the first-order
neighbourhood $FON(i)$ as the set of databases j with $A_{ij} > A_{th}$, and add
all databases of $FON(i)$ to $DB(\Pi_i)$.
Step 9: For all databases $j$ of $FON(i)$, select the database $k$ with
$A_{jk} > A_{th}$ and add database $k$ to the second-order neighbourhood $SON(i)$.
Step 10: If $A_{ik} > A_{th}$, add database $k$ to $DB(\Pi_i)$.
Step 11: Iterate step 9 and 10 for $m^{th}$ order-neighbourhood or until
there is no database left
Step 12: Return potential coalition $\Pi_i$ with $DB(\Pi_i)$
```

4.7.2 Management of Coalition Events

In the proposed approach, coalitions are managed with the coalition management operators that we present in this section. We propose that when database users reach an agreement on an action to be done (e.g., merge two coalitions), an administrator user can implement the changes with the appropriate operator and the change can be registered in a network management knowledge base. The effect of operators is to modify the structure of the ad hoc network. However, it is out of the scope of this thesis to indicate how users of databases can negociate such agreements. Also, we also assume that the users would verify that the coalitions created through these operators are viable – notably by making sure that their context is neither too general or too specific, which would make the resulting coalitions useless.

Create Coalition (Π, $DB(\Pi)$, *Att, t*) creates a new coalition Π with the set of databases $DB(\Pi)$. *Att* is the attractor database from which the coalition was created, and t is the time that the coalition was created. The context of the coalition $CTX(\Pi)$ is computed with the merging operators defined in section 4.5.

Excludes from Coalition (Π, $DB^{-}(\Pi)$, Π', *t*) creates a new coalition Π' by removing a set of databases $DB^{-}(\Pi)$ from an existing coalition Π at time *t*. A database may be excluded from a coalition when (1) the user's interest changes; (2) the context of the coalition changes because new databases are added, and the database's user interest are too dissimilar from the new context of the coalition.

Expand Coalition (Π, $DB^{+}(\Pi)$, Π', *t*) is the inverse operator of "exclude from coalition." It creates a new coalition Π' by adding a set of databases $DB^{+}(\Pi)$ to an existing coalition Π at time t. A coalition may expand when (1) a new database enters the network and the user's interest satisfies the constraints of the coalition; (2) the user's interest changes.

Merge Coalition (Π, Π^1, Π^2, *t*) is the operator to merge two existing coalitions Π^1 and Π^2. It creates a new coalition Π' by merging the sets of databases of Π^1 and Π^2 at time *t*. Two coalitions can merge if (1) their contexts are sufficiently similar; (2) their constraints are non-contradicting.

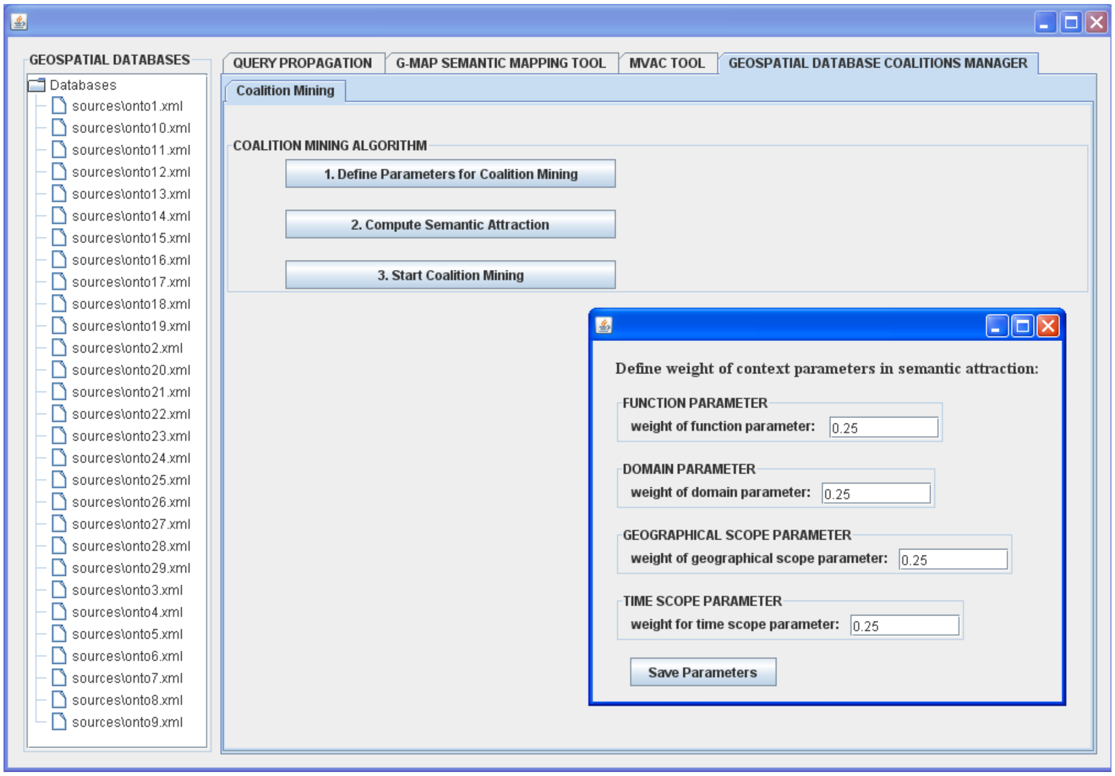
Divide Coalition (Π^1, Π^2, $DB^1(\Pi)$, $DB^2(\Pi)$, Π, *t*) is the inverse operator of Merge Coalition. It creates two coalition Π^1 and Π^2 composed of the sets of databases $DB^1(\Pi)$ and $DB^2(\Pi)$ respectively, with $DB^1(\Pi) \cup DB^2(\Pi) = DB(\Pi)$; *t* is the creation time for both new coalitions. A coalition can divide if it has become too large and its context can be divided into meaningful sub-contexts.

Dissolve Coalition (Π, *t*) is the operator that dissolves a whole coalition Π at time *t*. A coalition can dissolve when a large number of its members have quit the coalition and there are too few members to fulfill initial goals of the coalition.

4.8 Implementation and Evaluation of the Approach

Based on the proposed approach, we developed the Geo-Coalition Mining Tool for the discovery, the formation and the management of geospatial database coalitions. The Geo-Coalition Mining Tool is an interactive tool that guides the user in discovering the groups of databases that may be gathered and which users could use to collaborate toward a common goal. The Geo-Coalition Mining Tool implements the coalition-mining algorithm with the Java language on top of the JXTA peer-to-peer platform. JXTA simulates an open peer-to-peer network where autonomous peers can enter the network, send messages to each other, form groups, and quit the network. For a more detailed description of the JXTA platform, see Sun Microsystems, 2010. A formal representation of database context in the OWL-DL language was associated with each peer of the network. This choice of language is motivated by the reasoning capabilities of OWL-DL, and by the fact that OWL is the most popular language for formal ontologies in the Semantic Web. The databases' context were built from specifications of a set of geospatial databases, including among others the National Topographic Database of Canada (NTDB) (Natural Resources Canada, 1996), which contains information about roads, buildings, facilities, and waterbodies, and the International Disaster Database (EM-DAT, 1988). Globally, we used 30 geospatial databases during the experimentation, including databases storing data on topography, hydrography, disasters in Canada, meteorology, emergencies and demographics. The main interface of the tool accessible to any peer of the network is shown in Figure 4.6.

The tool allows the user to set the weights for the different context parameters. He or she can also set advanced parameters of the coalition-mining process, including the size of the network sample which gives the number of semantic attraction value computations at each cycle, the maximal number of iterations (cycles), the attraction threshold, and the threshold for the deviation of the degree of centrality. These parameters can be tightened or loosened when the coalition-mining algorithm finds too large coalitions, or finds no coalition, respectively. They influence the semantic aspect of coalitions, as more restrictive parameters will produce "more specialized coalitions" (more constraints) and less restrictive parameters will produce "more inclusive coalitions" (less constraints) in the semantic sense. For example, during the testing, the threshold for the deviation of the degree of centrality was finally set to 0.2. We have observed that with a higher deviation threshold, very few coalitions were detected for the set of sources being used. Below 0.20, more coalitions were detected, but the characteristics of their context were much more general (e.g., a coalition with any databases concerned with hydrography), which is less useful. This is illustrated in Figure 4.7.



Define weight of context parameters in semantic attraction:

FUNCTION PARAMETER

weight of function parameter: 0.25

DOMAIN PARAMETER

weight of domain parameter: 0.25

GEOGRAPHICAL SCOPE PARAMETER

weight of geographical scope parameter: 0.25

TIME SCOPE PARAMETER

weight for time scope parameter: 0.25

Figure 4.6 Determination of the user-defined parameters for coalition-mining process

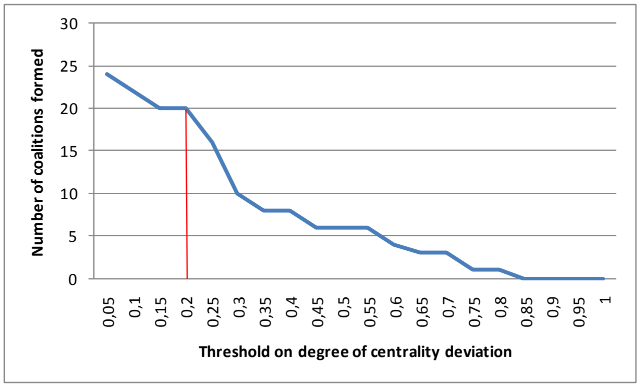
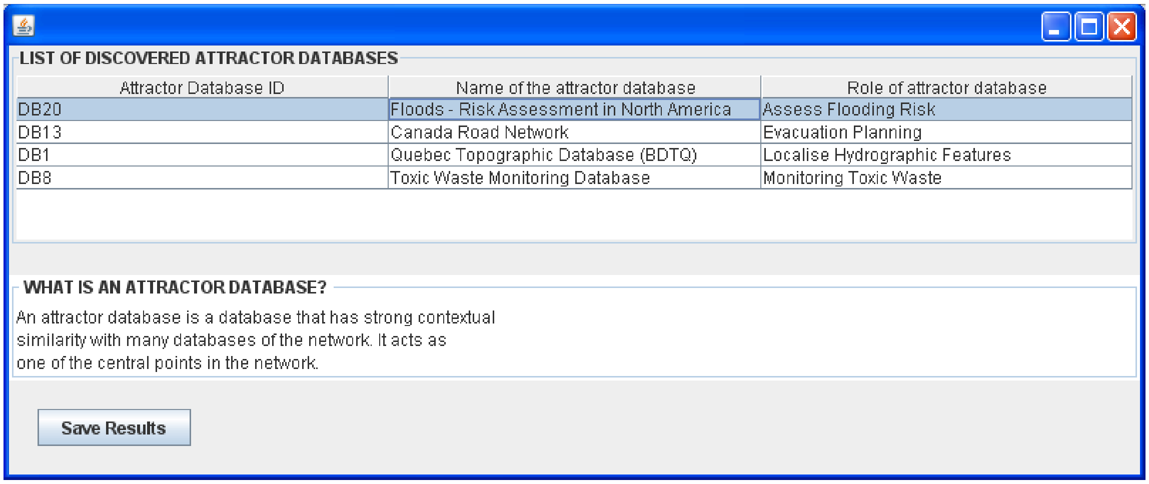


Figure 4.7 Experimentation on the number of coalitions found with various values of threshold for deviation of degree of centrality

The tool shows the attractor databases that were detected and the user can visualize the values of coalition-mining statistics, such as deviation of degrees of centrality, to change the coalition-mining parameters when he is not satisfied with the result. The list of discovered attractor databases is displayed with the function of the discovered attractor databases (for example, assess flooding risk or evacuation planning are the functions of two databases that were identified as attractors) (Figure 4.8).



LIST OF DISCOVERED ATTRACTOR DATABASES

Attractor Database ID	Name of the attractor database	Role of attractor database
DB20	Floods - Risk Assessment in North America	Assess Flooding Risk
DB13	Canada Road Network	Evacuation Planning
DB1	Quebec Topographic Database (BDTQ)	Localise Hydrographic Features
DB8	Toxic Waste Monitoring Database	Monitoring Toxic Waste

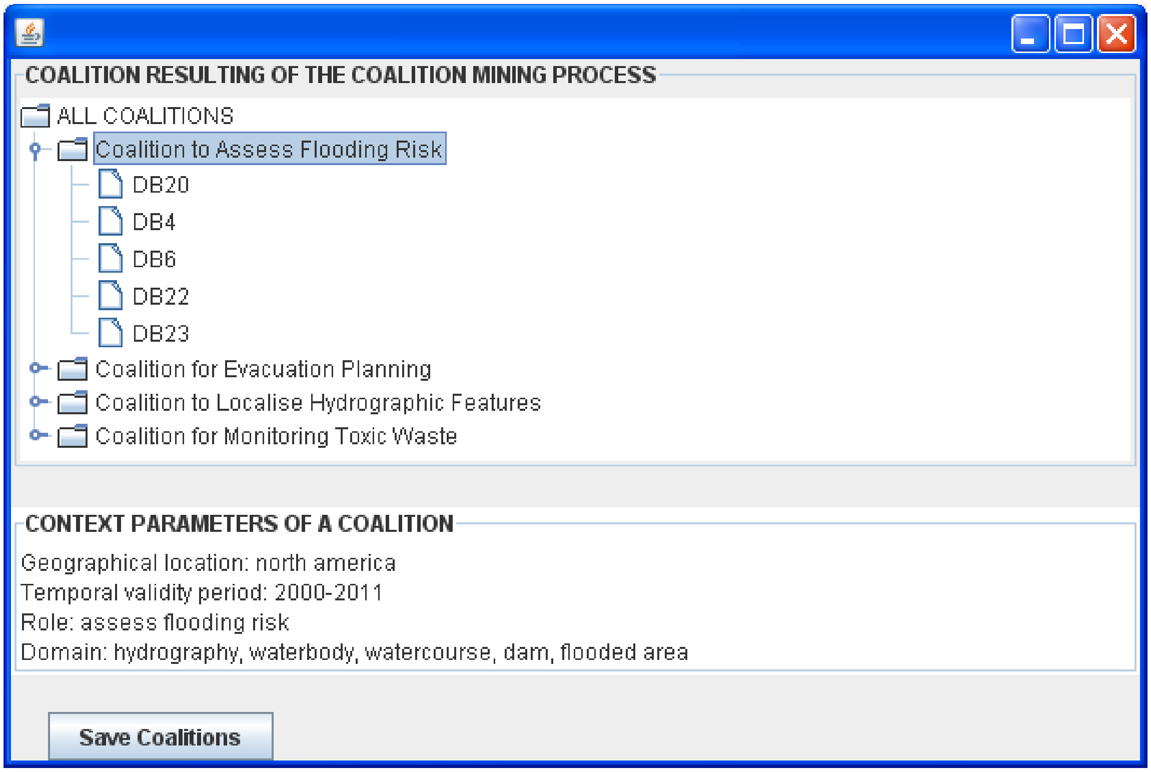
WHAT IS AN ATTRACTOR DATABASE?

An attractor database is a database that has strong contextual
similarity with many databases of the network. It acts as
one of the central points in the network.

Figure 4.8 Attractor databases are displayed to the user along with their function. The user can adjust the result by modifying coalition-mining parameters.

Then, the members of the coalition for each attractor database (the neighborhood of attractor databases) are identified. In Figure 4.9, the coalition browser allows visualizing the Coalition Tree, where geospatial databases coalitions that were discovered during coalition-mining process and created with the agreement of peers are classified according to a chosen context parameter. Here, they are classified according to their function, but they could be classified according to their domain, their geographical location, or their time period. The following shows an example of a coalition description (in OWL RDF/XML syntax). The description indicates the geographical location covered by spatial entities stored in the databases of the coalition (e.g., North America), the temporal validity period of the coalition, which corresponds to the time period during which the entities stored in the databases of the coalition were existing, as well as the function and domains of the coalition:

```
<Coalition rdf:ID="NorthAmericaFloodingRiskAssessmentCoalition"/>
<GeographicalLocation rdf:datatype="&xsd;string">NorthAmerica
</GeographicalLocation>
<TemporalValidityPeriodStart
rdf:datatype="&xsd;gYear">2000</TemporalValidityPeriodStart>
<TemporalValidityPeriodEnd
rdf:datatype="&xsd;gYear">2010</TemporalValidityPeriodEnd>
<Function rdf:datatype="&xsd;string"> AssessFloodingRisk</Function>
<Domain rdf:datatype="&xsd;string">
Hydrography</Domain>
<Domain rdf:datatype="&xsd;string">
Waterbody</Domain>
<Domain rdf:datatype="&xsd;string">
Watercourse</Domain>
<Domain rdf:datatype="&xsd;string">
Dam</Domain>
<Domain rdf:datatype="&xsd;string">
FloodedArea</Domain>
</Coalition>
```



COALITION RESULTING OF THE COALITION MINING PROCESS

ALL COALITIONS
- Coalition to Assess Flooding Risk
  - DB20
  - DB4
  - DB6
  - DB22
  - DB23
- Coalition for Evacuation Planning
- Coalition to Localise Hydrographic Features
- Coalition for Monitoring Toxic Waste

CONTEXT PARAMETERS OF A COALITION

Geographical location: north america
Temporal validity period: 2000-2011
Role: assess flooding risk
Domain: hydrography, waterbody, watercourse, dam, flooded area

Figure 4.9 Coalitions identified during the mining process are displayed in a tree

In order to measure the performance of the approach, we have assessed the number of meaningful coalitions that were retrieved, for different values of the semantic attraction threshold. By “meaningful” coalitions, we mean coalitions whose context is not too general, so it is not irrelevant. Figure 10 shows that the approach maintains, for a maximum number of 20 coalitions being formed, an acceptable number of meaningful coalitions until the threshold reaches 0,25. Below this threshold, we can expect that irrelevant databases are added to coalitions in the higher level neighbourhoods.

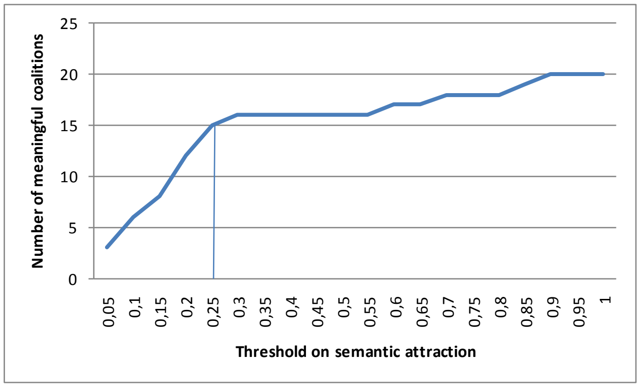


Figure 4.10 Experimentation on the meaningful coalitions found with respect to semantic attraction threshold

The purpose of coalitions is to support collaboration and partition the network according to various characteristics that facilitate the discovery of relevant sources. The contribution of this tool is the enhanced flexibility of coalition-mining with various criteria relevant to the geospatial domain, and the resolution of conflicts between the description of interests and contexts. In addition, because coalitions partition the network according to various context parameters, they can be used to support the propagation of queries to relevant sources of the network. Query propagation is the problem of determining, for a query issued by the user at a node of the network, to which databases this query should be forwarded, in order to optimize query results while accessing a minimal number of nodes. Query propagation could beneficiate from coalitions, since the later could define the scope of the query.

4.9 Conclusion and Future Work

Coalition of geospatial databases in ad hoc networks are essential to gain from the interconnected world and to facilitate meaningful geospatial data sharing and collaboration. In this paper, we have proposed a conceptual coalition model to support the discovery of coalitions among databases of an ad hoc network. In addition to this model, we have defined the concept of semantic attraction among databases and among coalitions of databases. The concept of semantic attraction supports the automatic discovery of databases that have affinities because they want to be part of similar coalitions from a thematic, spatial and temporal point of view, or because they have complementary knowledge from those three perspectives as well. Finally the concept of semantic attraction is integrated into a coalition mining algorithm, which uses network analysis techniques to discover potential coalitions. There are still many open issues that were raised in this paper and that will be addressed in future work. In particular, we plan to integrate this approach in a global framework on real time semantic interoperability in ad hoc networks of geospatial databases, to show the impact of coalition in this context. In this future work, it will be investigated how coalitions in ad hoc networks can support the propagation of queries to relevant databases.

Acknowledgments

The author wishes to acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

CHAPTER 5

Semantic Augmentation of Geospatial Concepts: The Multi-view Augmented Concept to Improve Semantic Interoperability between Multiple Geospatial Databases

M. BAKILLAH, M.A. MOSTAFAVI

Published in Bo Wu (eds.) ISPRS Book, Taylor & Francis, 2011. Extended version was published in *Journal of Earth and Engineering*.

5.1 Présentation de l'article

Le prochain composant du cadre conceptuel pour l'interopérabilité sémantique en temps réel qui est présenté dans ce chapitre est consacré au problème de la représentation des connaissances. En se basant sur le modèle du Concept multi-vues augmenté présenté au Chapitre 3, nous présentons une méthode pour l'extraction et la génération de concepts MVAC, laquelle sera publié dans la conférence *Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science*, Hong Kong, 26-28 mai 2010.

L'interopérabilité sémantique est une problématique cruciale qui doit être résolue pour assurer l'échange de données géospatiales entre les bases de données géospatiales, ainsi que l'interprétation de ces données. Afin de réaliser l'interopérabilité sémantique, des mappings sémantiques doivent être établis entre les concepts faisant partie des ontolgies qui décrivent ces bases de données. Cependant, les mappings sémantiques ne peuvent être établis que si la sémantique des concepts est explicite. Les définitions existantes des concepts ne sont pas toujours suffisantes pour représenter toute la richesse et tous les aspects des concepts géospatiaux. De plus, la sémantique peut être implicite, faisant en sorte qu'elle ne peut être exploitée pendant le processus de mapping sémantique. Cet article présente une nouvelle représentation des concepts géospatiaux, soit le Concept multi-vues augmenté (MVAC), laquelle prend en compte les problèmes identifiés ci-dessus. Ensuite, nous proposons une méthode pour générer les concepts MVAC, laquelle comprend : (1)

une méthode pour l'extraction des différentes vues d'un concept qui sont valides dans différents contextes, puis (2) une méthode d'augmentation du concept qui ajoute au concept les dépendances implicites qui peuvent exister entre ses caractéristiques, dépendances qui peuvent être extraites au moyen de techniques de fouille de règles d'association. L'approche proposée jouera un rôle important pour améliorer la qualité de l'interopérabilité sémantique entre les bases de données géospatiales, puisqu'elle prend en compte les relations sémantiques implicites entre différents concepts.

5.2 Presentation of the Article

The next component of the real time semantic interoperability framework addresses the issue of knowledge representation. Based on the Multi-View Augmented Model presented in Chapter 3, a method for extraction and generation of MVAC concepts is presented in this chapter, which will be published in the *Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science*, Hong Kong, 26-28 May 2010.

Semantic interoperability is a key issue for the meaningful sharing of geospatial data between multiple geospatial databases. It requires that semantic mappings can be established between concepts of ontologies of those databases. Semantic mappings can be discovered only when semantics of concepts are explicit. However, existing definitions of concepts are not always sufficient to represent all the semantic richness of geospatial concepts. In addition, some semantics may be implicit in the ontologies, and implicit semantics cannot be used during the semantic mapping process. This paper proposes a new representation for geospatial concepts, called the Multi-View Augmented Concept (MVAC), which takes into account these drawbacks. Next, we propose a method to generate a MVAC, based on: (1) extraction of the different views of a concept that are valid in different contexts, and (2) augmentation of a concept with implicit dependencies between its features based on rule mining theory. We believe that the proposed approach will play an important role to improve the quality of the semantic interoperability between multiple geospatial databases since it takes into account the implicit semantic relations between different concepts.

5.3 Introduction

Semantic interoperability is a major research topic for ensuring data sharing and reuse among heterogeneous systems (Bian and Hu 2007). It is the knowledge-level interoperability that provides cooperating databases with the ability to resolve differences in meanings of concepts (Park and Ram 2004). To resolve differences in the meaning of concepts, this meaning must be made available to machines through an explicit representation; by doing so, differences can automatically be identified during the semantic mapping process. However, current semantic mapping approaches rely on poor definition of concepts that is not suitable for representing all the semantic richness of a geospatial concept. For example, not considering explicitly the semantics of spatial and temporal properties of a geospatial concept reduces its expressivity and semantic richness. In addition, it may contain implicit knowledge that was not explicitly represented but that can be inferred from existing knowledge. The structure of the concepts is also valuable in defining its semantics. This is why considering a concept as a bag of features is not sufficient. To address these problems, we propose a new representation of geospatial concepts, called the Multi-View Augmented Concept Model (MVAC) (presented in section 3), and a method to generate MVAC representations from the original concepts being defined in an ontology (presented in section 4). In this method, we add two additional layers to the definition of the concept. First, we extract the different views it can have in different contexts, and then, we augment it with new dependencies between its features. The contribution of the MVAC model is to improve semantic interoperability with a concept that has richer semantics, and a structure that supports thediscovery of semantic relations between concepts of different ontologies, relations that were hard to discover with traditional, lexical-based semantic mapping approaches.

This chapter is organized as follows: in section 5.4, we review related work on the definition of concepts. In section 5.5, we propose the MVAC model. In section 5.6 we propose the MVAC generation method. In section 5.7 we discuss with a case study how the MVAC can help to improve semantic interoperability. In section 5.8 we conclude this paper.

5.4 Related Work

Knowledge representation is the problem of encoding the knowledge that human have about reality, in such a way that it supports reasoning (Kavouras and Kokla 2008). A knowledge representation is not a complete and perfect picture of the reality; but an imperfect abstraction of a portion of reality that is relevant in an application domain. Knowledge representation is a fundamental issue for improving semantic interoperability because it supports knowledge sharing (between humans and between machines). The theoretical basis of knowledge representation approaches depends on the different theories of concept. From a cognitive point of view, concepts are mental representations of a category (Medin and Rips 2005), and a category denotes a set of real world entities that have similar properties, relations, functions, etc. (Kavouras and Kokla 2008). It is very difficult to give a framework that would guide the assignment of properties to concepts in a universal way, even if such attempts were made (Margolis and Laurence 1999; Bennett 2005). Mostly, the choice of the properties of a concept depends on the purpose or intended task (Tomai and Kavouras 2004). In the geospatial domain, existing definitions of concept focus on the identification of the special properties of geospatial objects. According to Kavouras and Kokla (2008), a concept can be represented with a term, a set of semantic elements (properties and relations) and their values. This is similar to the concept definition proposed by Schwering and Raubal (2005), where concept is defined with properties (represented as dimensions in a conceptual space) and their values (represented as values of these dimensions). The following properties and relations were mentioned by Kavouras and Kokla: *purpose*; *agent*; *shape*, *size*, *cover*, *property-defined location*; *frequency*, *duration*, *property-defined time*; *is-a*, *part-of* relations; *relative position relations* (upward, downward, behind, etc.); *proximity*, *direction* and *topological relations* (adjacency, connectivity, overlap, etc.); and *source-destination relation*. In their concept definition, Rodriguez and Egenhofer (2003) have classified features of concept as attributes, functions (representing what is done to or with an object) and parts (structural component of an object). This classification of properties aims at facilitating the separate manipulation of each type of property, more specifically in the context of semantic similarity assessment. Another set-based definition of concept is given by Brodeur and Bédard (2001). They have proposed a concept definition based on the four-intersection model of Egenhofer (1993). Concepts have an interior, defined by their intrinsic properties (e.g. identification, attributes, attribute values, geometries, temporalities, and domain), and a boundary, defined by their extrinsic properties (e.g., semantic, spatial, and temporal relationships and behaviours). According to this definition, the whole set of intrinsic and extrinsic properties forms the *context*. However, Keßler et al. (2007) argue that the context has two components: the internal context specifies the domain of application and the external context is defined by a set of rules that allows modifying the concept in different circumstances.

These definitions of concept are mainly targeted at the geospatial domain. Bennett (2005) has attempted to provide a generic concept definition. He proposes that properties of an object may be classified as physical (including geometry and material properties); historical (how the object came into existence; the events it has undergone, etc.); functional, including static and dynamic functions; and conventional properties (related to the *fiat* nature of objects). Although Bennett mentions that "objects that exhibit one property, will very often also exhibit another property", he does not explicit further those types of dependencies between properties. A first problem with the above-mentioned approaches is that they define concept as an unstructured set of features. However, features are related to each other through dependencies, which are considered as improving the expressivity of ontologies (Curé and Jeansoulin 2007). For example, the position of a moving object depends on time, the value of an object's temperature depends on its altitude, etc. The concept definition cannot be complete if we do not represent these dependencies. However, if these dependencies are not stated in a given concept's definition, it may be possible to discover implicit dependencies by looking in the instances of the concept.

A second problem is that in most of the definitions, spatial and temporal properties are not explicit but merged into other classes of properties. This means that the separate manipulation of spatial or temporal properties is difficult because they are not explicit and cannot be efficiently used for semantic interoperability purposes. Also, most approaches define properties only with their name and range of values, for example the "geometry of a house" is a "polygon". But this is not sufficient to understand the exact semantics of this spatial property. This is because the polygon may represent the "roof of the house" or the "foundation of the house". Therefore, spatial and temporal properties have to be further described in a more explicit manner.

Finally, there are different ways to define a given concept depending on the context (Parent *et al.* 2006). Hence, several researchers have recognized the existence of a multi-view paradigm for concepts and propose modelling views in geospatial databases (Bédard and Bernier 2002; Parent et al. 2006) and in ontologies (Bhatt et al. 2006; Wouters et al. 2008). Besides the strict representation issues, multiple views of a concept can also provide multiple ways to achieve semantic interoperability. However, existing representations of geospatial concepts usually do not address this paradigm explicitly, nor demonstrate its usefulness in semantic interoperability.

5.5 The Multi-View Augmented Concept Model (MVAC)

This new concept definition that we propose is intended to address the drawbacks of the above-mentioned concept definitions. Its contribution, with respect to exiting concept definitions, is to provide a richer and more structured concept definition as a basis to improve semantic interoperability. As explained in Section 3.6.2, the MVAC adds two additional levels of definition to the original definition of the concept: a set of views valid in different contexts, and a set of dependencies between features of the concept.

At the basic level of definition, a concept, denoted by c, is defined as: c = <n(c), {p(rp)}, {r(rr)}, {spatial_d(rsd)}, {temporal_d(rtd)}>, where:

- n(c) is the name of the concept;
- {p(rp)} is the set of properties of the concept. The set of possible values of a property, called the range and denoted rp, is given in brackets.
- {r(rr)} is the set of relations that c has with other concepts. rr represents the range of the relation r, that is, the set of concepts c is linked with through relation r.

- {spatial_d(rsd)} is a set of properties, called spatial descriptors, whose role is to describe the spatiality of the concept. For example, the concept watercourse could have the spatial descriptor geo-entity (axis of watercourse), meaning that the line geometry representing the watercourse corresponds to the axis of the watercourse. The range of a spatial descriptor is denoted rsd.

- {temporal_d(rtd)} is a set of properties, called temporal descriptors, whose role is to describe the temporality of the concept. The range of temporal descriptors is denoted rtd. For example, the concept watercourse may have temporal descriptor waterlogged period(average flooded period) which means that the waterlogged period corresponds to the average time the watercourse is flooded over years. We give an example of a concept "watercourse":

c = <watercourse, {water level(low, medium, high), category(intermittent, stable), spatial extent(polygon, moving polygon), function(navigation, skating, evacuation area), state(frozen, unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, flooded area, frozen area)}{waterlogged period(average flooding period)}>

Such concept may represent different realities in different contexts. For each context, we want to create a view that could be used in that context. In a published paper (Bakillah et al. 2009), we have stated that the view paradigm supports ontology reuse, by selecting only parts of a concept that are relevant in a given context. A view of a concept is therefore a selection of its features that are valid in a given context. A view of a concept c is represented as:

View(c): Context(Name of context) → <{p(rp_v)}, {r(rr_v)}, {spatial_d(rsd_v)}, {temporal_d(rtd_v)}>

This expression means that in the identified context, the set of values for a property, or the range of a relation or of a descriptor is a subset of the original set of values or range, denoted rp_v, rr_v and rsd_v, rtd_v respectively. For example, two possible views of the concept watercourse may be the following (the features whose values are affected are in bold):

Context(flooding) → <{**water level**(high), category(intermittent, stable), **spatial extent**(moving polygon), **function**(evacuation area), **state**(unfrozen)}, {Connect(Waterbody)}, {**geo-entity**(flooded area)}{waterlogged period(average flooding period)}>

Context(tourism) → <{water level(low, medium, high), category(intermittent, stable), **spatial extent**(polygon), **function**(navigation, skating), state(frozen, unfrozen)}, {Connect(Waterbody)}, {**geo-entity**(bed of watercourse, frozen area)}{waterlogged period(average flooding period)}>

The example shows that in the context of a flood, the value of water level is necessarily high, the watercourse could have the function of an evacuation channel that would be used by boats to rescue people, and its spatial extent would be represented by a moving polygon to represent the evolution of the flooded area. A view is a spatial view when the condition is imposed only on a spatial property, a spatial relation (topology, proximity, and orientation) or a spatial descriptor:

Spatial View: Context(Name of context) → spatial property (concept, value of spatial property)

Spatial View: Context(Name of context) → spatial relation (concept, range of spatial relation)

Spatial View: Context(Name of context) → spatial descriptor (concept, value of spatial descriptor)

Similarly, a view is a temporal view when the condition is imposed on a temporal property, a temporal relation or a temporal descriptor:

Temporal View: Context(Name of context) → temporal property (concept, value of temporal property)

Temporal View: Context(Name of context) → temporal relation (concept, range of temporal relation)

Temporal View: Context(Name of context) → temporal descriptor (concept, value of temporal descriptor)

At the augmentation level, dependencies between features can be inferred to semantically augment a concept. Dependencies express that a first feature's values are constrained by a second feature's values. We formalize dependencies with rules, in the form: head → body. The body in the rule is a consequence of the head. Here are examples of thematic, spatial and temporal rules respectively:

Altitude(land, low)→ FloodingRisk(land, high)

Width(watercourse, larger than 7m)→ Geometry(surface)

Flooding frequency(land, more than twice a year)→ Status(land, periodically waterlogged).

The concept, plus the views and the augmented dependencies, forms the multi-view augmented concept (MVAC), which is defined as follow:

$c^{MVA} = <$ n(c), {p(c)}, {r(c)}, {spatial_d(c)}, {temporal_d(c)}, {v(c)}, {dep(c)}>

where {v(c)} is the set of views, and {dep(c)} is the set of augmented dependencies. The methodology that will augment a concept to a MVAC will therefore be composed of three main methods: a context extraction method, a view extraction method, and a method to discover dependencies between features.

5.6 Context Extraction

The aim of context extraction is to identify the context(s) of a concept, in order to facilitate the interpretation of the meaning of this concept and extract views of the concept that are valid in each context. Most existing context extraction approaches were developed in support of context-aware systems (Baldauf et al. 2007). Generally, context models that formalize common elements of context (such as user profile, localisation, time, tasks, preferences, etc.) are proposed, and several sensors that monitor the environment of a user or a device are responsible for gathering the values of the context element. We follow a similar approach, that is, we propose a context model that formalizes the main elements of the context of a concept. However, the information source for identifying the context(s) of a particular concept is necessarily different. With respect to concepts, the main sources of information that are helpful to identify the context of concepts are the set of database specifications, which contain the definitions of concepts. For example, a lake can be defined as "a body of water surrounded by land." Figure 5.1 illustrates the context model with UML schema. The development of the model was based on the systematic analysis of existing geospatial database specifications to identify types of contexts that characterize concepts.

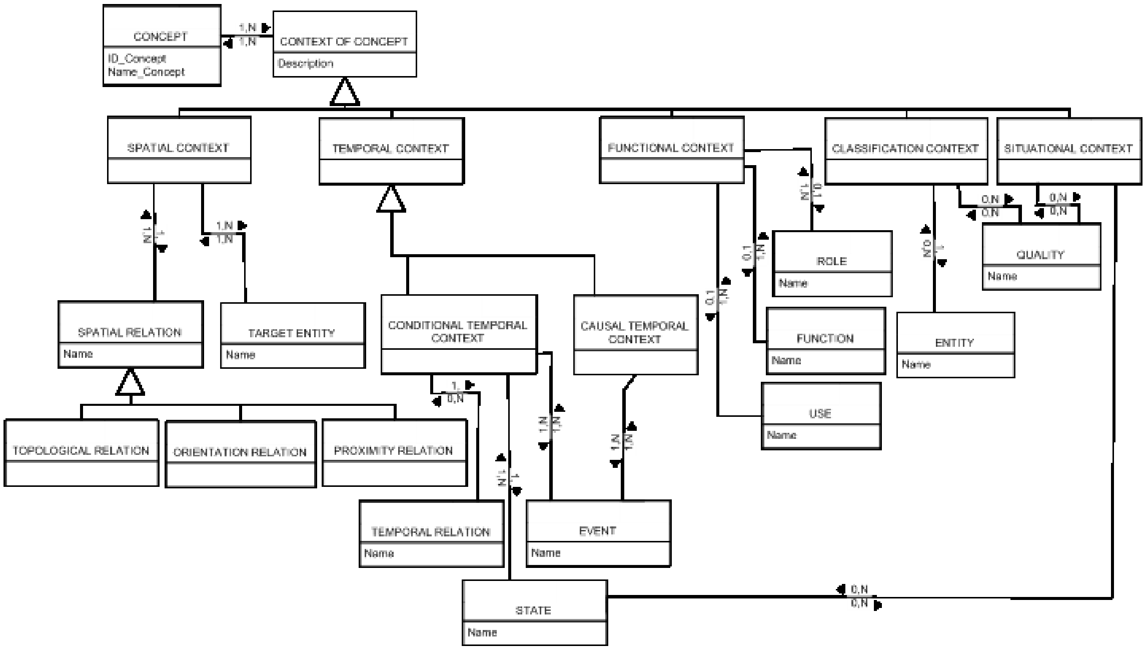


Figure 5.1 Context Model

We followed a bottom-up approach, where we have manually identified the elements that participate in the definition of concepts; then we have classified them into five categories of contexts:

- **functional context**: a function, task or role associated with the concept, for example, *navigation* is a possible functional context of the concept *water canal*. Functional context can be identified with expressions such as: *used for*, *used to*, *purpose*, *for*, etc.
- **situational context**: something that describes the situation, state of quality of entities represented by the concept, for example, *dryness* is a possible situational context of the concept *land*. Adjectives are useful to identify situational context.
- **classification context**: a category of things to which the concept belongs; for example, *bridge* can have as classification context *transport infrastructure* or *hazard to air navigation*. The classification context is often identified by the genus of the concept's definition. The genus is the super-ordinate term in the definition. For instance, in "*brook: natural stream of water smaller than a river*", "*stream of water*" is the genus of "*brook*". The classification context can also be identified with expressions such as *type of* or *is a kind of*.
- **spatial context**: the description of a geographic area or place, that may be described with a spatial relation of proximity, topology, or orientation (for example, *close to floodplain*, *surrounded by land*);
- **temporal context**: a period of time, or an event, activity or change (for example, an historical period or the event of *earthquake*). Temporal context may be of two kinds: a temporal relation of the concept with an event (ex: *after an earthquake*) or a conditional temporal context, where a state or a quality of entities represented by the concept is temporally dependant on the occurrence of an event (ex: *a stream where debit depends on precipitations*).

The proposed approach is based on the idea that concepts' definitions implicitly contain elements that describe context. By analysing definitions, our aim is to identify implicit context elements and represent them explicitly. The proposed extraction method is based on the detection of *context patterns* in definitions. Context patterns are generic and relatively short expressions that are often used in definitions to describe context elements. Although the objective of this research is not to provide an exhaustive set of context patterns, Table 5.1 presents some context patterns that were identified from definitions of geospatial concepts. After each pattern, we provide the formalized context expression that is created when the corresponding pattern is identified in a concept’s definition. The oblique / is used to separate different elements of a pattern (e.g., the pattern *“resulting from”/article/name OR composed name* means that the expression *“resulting from”* is directly followed by an *article*, which is directly followed by a *name* or a *composed name*). In several cases, the extracted context expression is the same, but semantically equivalent patterns are transformed into a single expression for homogeneity. For example, *caused by* and *resulting from* are both expressed as *caused by*.

Table 5.1 Context patterns for geospatial concepts

Type of context pattern	**Pattern**	**Extracted context expression**	**Example**
Spatial context patterns	Spatial relation/article OR “of”/name OR composed name Spatial relation/ name OR composed name “where”/proposition	Spatial relation/article OR “of”/name OR composed name Spatial relation/ name OR composed name “situated where”/proposition	On top of the roof Surrounded by water Beside apartment buildings Where water level increases
Temporal context patterns	Name/ “depends on” OR “caused by” /name OR composed name Adjective/”when” /proposition “resulting from”/ name OR composed name “resulting from”/ article/name OR composed name Temporal relation/ article/name OR composed name Temporal relation/ name OR composed name	Name/ “depends on” OR “caused by” /name OR composed name Adjective/”when” /proposition “caused by”/ name OR composed name “caused by”/ article/name OR composed name Temporal relation/ article/name OR composed name Temporal relation/ name OR composed name	Debit depends on precipitations Is caused by a storm Flooded when precipitations are important Resulting from landslide Resulting from the exploitation After an earthquake
Functional context patterns	Verb/article/name OR composed name Verb/name OR composed name “ used for”/name OR composed name “used for”/article/name OR composed name “ used for”/verb	Verb/article/name OR composed name Verb/name OR composed name “ used for”/name OR composed name “used for”/article/name OR composed name	Drain an area Stores cereals Used for navigation

	“is used for”/article/verb “ used to”/verb/name OR composed name “ used to”/name OR composed name “used to”/article/name OR composed name “ used to”/verb “used to”/article/verb “ for the purpose of”/name OR composed name “for the purpose of”/article/name OR composed name “for the purpose of”/verb “for the purpose of”/article/verb “ for”/name OR composed name “for”/article/name OR composed name “for”/verb “for”/article/verb "for" AND article AND proposition “build to” OR “build for” OR “designed to”/proposition “that allows”/ article/name OR composed name “that allows to”/ proposition	“ used for”/verb “is used for”/article/verb “ used to”/verb/name OR composed name “ used to”/name OR composed name “used to”/article/name OR composed name “ used to”/verb “used to”/article/verb “ used for”/name OR composed name “used for”/article/name OR composed name “for”/verb “for”/article/verb “ for”/name OR composed name “for”/article/name OR composed name “for”/verb “for”/article/verb "for" AND article AND proposition “for” OR “designed to”/proposition “for”/ article/name OR composed name “for”/ proposition	Used to produce electricity For the purpose of transportation For navigation For the displacement of vehicles Build for the storage of material That allows the movement of vehicles
Situational context patterns	Adjective/genus Adjective/name OR composed name/genus “that can be” OR “which can be”/verb Adverb/adjective concept AND ("in" OR "within") AND adverb adjective AND concept	Adjective Adjective/name OR composed name “can be”/verb Adverb/adjective ("in" OR "within") AND adverb adjective	Small Low level that can be displaced usually dry often waterlogged in construction abandoned roads
Classification context patterns	Genus of the definition “part of“/article/name OR composed name “part of“/name OR composed name “which is” /article/name OR composed name Adjective/Genus of the	Genus of the definition “part of“/article/name OR composed name “part of“/name OR composed name name OR composed name Adjective/Genus of the	Canal : artificial **water path**. Part of a building Which is an hazard for navigation Flooded land

	definition	definition	
	“portion of“/article/name OR composed name	“portion of“/article/name OR composed name	Portion of a road
	“portion of“/name OR composed name	“portion of“/name OR composed name	
	“kind of”/name OR composed name	name OR composed name	Kind of human-made structure
	“type of”/name OR composed name	name OR composed name	

5.7 MVAC Generation Method

We have developed this method to transform a concept into a MVAC. The method integrates a view extraction paradigm, techniques for mining rules and principles for ontological reasoning. The approach for generating an MVA ontology (taking as input an original ontology) is depicted in Figure 5.2. It consists of two main phases: 1) the *view extraction* phase, and 2) the *augmentation* phase, which correspond to the additional levels of the concept's definition presented in Section 5.5. The method takes as input an ontology with original concepts as defined in Section 5.5. The first step involves the specification of the extraction rules by the user.

Step 1. Extraction of context and determination of view extraction rules. The specification of extraction rules allows for the interaction between users and the view extraction algorithm. The users specify what are the values of the properties, relations and descriptors of a concept that are valid in some context. For example, considering the concept “watercourse” with properties “depth” and “category of watercourse”, the user could declare some extraction rules specifying the possible values of those properties in the context of dryness:

Context(dryness) → water level(watercourse, low) (rule 1)

Context(dryness)→ category of watercourse (watercourse, intermittent) (rule 2)

Step 2. Inference of new extraction rules. Now that we have a set of extraction rules for the context of the concept, we want to check if new extraction rules can be inferred by combining them. We also use other existing rules that are already part of the ontology, and which represent the knowledge of domain experts. This is a way of reusing the existing knowledge to produce new knowledge. The inference of new extraction rules is a process that (1) takes as input the extraction rules that were specified in step 1, plus the rules that are already part of the ontology, (2) sends these rules to an inference mechanism, (3) produces new inferred rules, and (4) restarts the cycle from (1) to (3) until no new rules are inferred.

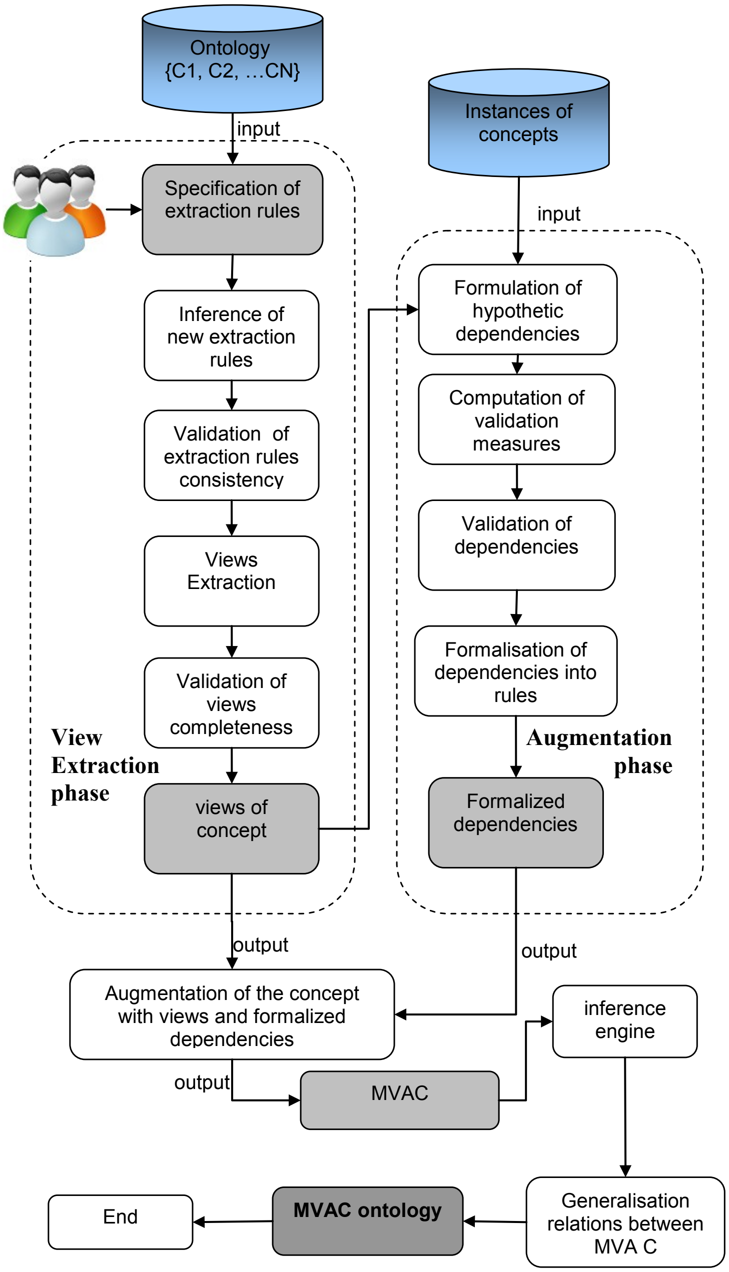

Ontology
{C1, C2, …CN}
input
Specification of extraction rules
Inference of new extraction rules
Validation of extraction rules consistency
Views Extraction
Validation of views completeness
View Extraction phase
views of concept
Instances of concepts
input
Formulation of hypothetic dependencies
Computation of validation measures
Validation of dependencies
Formalisation of dependencies into rules
Augmentation phase
Formalized dependencies
output
output
Augmentation of the concept with views and formalized dependencies
output
MVAC
inference engine
Generalisation relations between MVA C
MVAC ontology
End

Figure 5.2 MVAC and Ontology Generation Method

The inference mechanism verifies whether the body of a rule implies the head of a second rule; if so, the head of the first rule implies the body of the second rule. For example, consider a rule of the ontology stating that intermittent watercourses are represented by a moving polygon:

Category of watercourse(watercourse, intermittent) → geometry(watercourse, moving polygon)

From this rule and the ones that were specified by the user in step 1, we can infer the following new rule:

Context(dryness) → geometry(watercourse, moving polygon) (rule 3)

New inferred rules are added to the set of rules that will be used to extract views of the concept.

Step 3. Validation of extraction rule consistency. Before using these rules to extract the views of a concept, we need to verify if the rules that were inferred are correct, that is, if they are consistent with reality. In this case, reality corresponds to the instances of the concept, which are representations of real world objects stored in the database. To verify if the rules are accurate, we will assess the consistency between the rule and the instances. Consistency can be defined as the degree of consistency of the data with respect to its specifications (Mostafavi et al. 2004). In our context, data corresponds to instances whereas specifications correspond to rules (since rules define the semantics). Therefore, a rule is said to be consistent if the instances of the concept verify this rule. For example, if we have a rule Context(dryness) → water level(watercourse, low), we need to check if instances of the concept "watercourse" which have for value of the context "dryness", also have "low water level". To determine whether an extraction rule is consistent enough, we propose a ratio that will compare the number of instances that respect the rule (denoted with |verifying instances|) with the total number of instances which have for context the one indicated in the rule (denoted with |targeted instances|):

$$\text{Degree of consistency } = \frac{|\text{verifying instances}|}{|\text{targeted instances}|}$$

Only those rules that have a sufficient degree of consistency can be used for view extraction.

Step 4. View extraction. View extraction, as we have defined in previous work (Bakillah et al. 2009), includes two main steps, the extraction of partial views and the merging of partial views. First, we define the partial view as follows: *A partial view of a concept C is a sub-concept of C whose definition is constrained by the consequent of a single view extraction rule; that is, a partial view of C is the result of applying a single extraction rule on C*. The extraction of partial views means that each extraction rule is applied to the concept to create the subconcept that will always respect this rule. For example, for the concept watercourse defined in section 3, applying rule 1 gives the following partial view:

Partial view: Context(dryness) → <watercourse, {water level(low), category(intermittent, stable), spatial extent(polygon, moving polygon), function(navigation, skating, evacuation area), state(frozen, unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, flooded area, frozen area)}{waterlogged period(average flooding period)}>.

This partial view imposes a restriction only on the values of property "water level." In the second step of the view extraction, all partial views that pertain to a same context and that are non-contradicting are merged into a single view. This is the partial view merging process. For example, merging partial views generated by rule 1 to 3 would lead:

view: Context(dryness) → <watercourse, {water level(low), category(intermittent), spatial extent(moving polygon), function(navigation, skating, evacuation area), state(frozen, unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, flooded area, frozen area)}{waterlogged period(average flooding period)}>

During the view extraction, relations between views of a concept and other concepts of the ontology are inherited from the definition of the concept when it applies, for example the above view is linked to the concept "waterbody" with the spatial relation "connect."

Step 5. Validation of view completeness. When all views of a concept are created, we check if they are complete. This means, in our context, that the union of all views of the concept result in the concept itself. Remember that the restricted range of a property p_i (or relation R_i, descriptor d_i) in a view v_j is r_{ij}. The view completeness can be validated if the following generic expression is verified:

$c = < n(c),\ \{p_1(r_{11} \cup r_{12} \cup r_{13} ...), ... p_n(r_{n1} \cup r_{n2} \cup r_{n3} ...) \}, \{R_1(r_{11} \cup r_{12} \cup r_{13} ...), ... R_n(r_{n1} \cup$
$r_{n2} \cup r_{n3} ...) \},$
$\{d_1(r_{11} \cup r_{12} \cup r_{13} ...), ... d_n(r_{n1} \cup r_{n2} \cup r_{n3} ...) \}>.$

That is, by taking, for all features of the concept, the union operator on the restricted ranges of all views of the concept. The next steps are about augmenting the concept (with its views) with implicit dependencies.

Step 6. Formulation of possible dependencies. Possible dependencies are dependencies that have to be verified against data. For every view of a concept, our method formulates dependencies that express relations between each pair of their features (properties, relations or descriptors). Those dependencies are expressed as rules. For example, for a concept "watercourse" with properties "state (frozen, unfrozen)" and "function(skating, navigable)", we can have:

"If *state* of watercourse = frozen, then *function* = skating"
"If *state* of watercourse = frozen, then *function* = navigable"
"If *state* of watercourse= unfrozen, then *function* = skating"
"If *state* of watercourse= unfrozen, then *function* = navigable"

"If *function* of watercourse = skating then *state* = frozen"
"If *function* of watercourse = skating then *state* = unfrozen"
"If *function* of watercourse = navigable then *state* = frozen"
"If *function* of watercourse = navigable then *state* = unfrozen"

Because the number of possible dependencies may be high, they can be classified (the first series being classified as “function depends on state” rules, and the second as “state depends on function” rules) so that the user can reject the ones that seems non verifiable.

Once we have formulated a set of possible dependencies, we have to validate which ones are true among instances of a view.

Step 7. Computation of rule validation measures. For each rule expressing a possible dependency, we determine the values of two measures that will help to determine if we can retain it as a valid dependency. Those measures, which are *support* and *confidence*, are adapted from the domain of rule mining, which aim at finding correlations between items in datasets (Ceglar and Roddick, 2006). The support measure indicates how many instances respect either the head (Ihead) or the body(Ibody) of a rule, with respect to the total set of instances (Itotal):

$$Support = \frac{|Ihead \cup Ibody|}{|Itotal|}$$

The confidence measures how many instances respect the body of the rule among those that respect the head of the rule:

$$Confidence = \frac{|Ibody|}{|Ihead|}$$

Therefore, we define the valid dependency as follows: a valid dependency of a view *v* is a dependency for which the set of instances of view *v* respect the support and confidence threshold. Since these thresholds can be set by the user, the validity of a dependency may depend on the level of tolerance required by the user.

Step 8. Validation of dependencies. For the validation of dependencies, we choose those dependencies for which support and confidence values are reaching a pre-established threshold. Those measures complete each other since a high confidence but a low support means while this rule is usually respected, it is not really frequent in the instance set, so it may be less interesting.

Step 9. Formulation of dependencies into rules. If the rule checked in step 4 is determined to be true, then it is added to the definition of the view in a form: Feature 1(concept, value of feature 1) → Feature 2(concept, value of feature 2). With respect to technologies, we note that the extraction of dependencies can be supported by reasoning tools such as the Jena reasoner. Jena is a tool that implements an instance-based reasoner for OWL ontologies.

Now that views and dependencies are extracted, the concept's definition is rewritten with those new elements. However, relations between views and augmented concepts need to be re-computed to form the MVA ontology.

Step 10. The inference of relations. Views need to be linked together by generalisation/specialisation relations in order to create the MVA ontology. These links needs to be established between the different views of a same concept, and between views of different concepts. Generalisation is when the instances of a first view /concept include all instances of a second view/concept. To perform this task, we can express MVACs with the OWL-DL language and use a subsumption reasoning mechanism provided by a reasoning engine. For example, view 1 is a generalization of view 2:

View1: Context(dryness) → <watercourse, {water level(low), category(intermittent), spatial extent(moving polygon), function(non navigable, skating), state(frozen, unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, frozen area)}{waterlogged period(average flooding period)}>

view2: Context(dryness in summer) → <watercourse, {water level(low), category(intermittent), spatial extent(moving polygon), function(non navigable), state(unfrozen)}, {Connect(Waterbody)}, {geo-entity(bed of watercourse)}{waterlogged period(average flooding period)}>

This means that view 2 represents a smaller number or real world objects than view 1, and all instances of view 2 are instances of view 1. Therefore, views can be categorised within the MVA ontology.

5.8 Case Study

Having defined the MVA model and a method to generate it from an existing concept, we aim to show with the following examples that the MVAC can help to improve semantic interoperability. Consider the user of a geospatial database whose ontology contains the concept "watercourse," defined as follows:

C1: <**watercourse**, {water level(low, high), spatial extent(polygon, moving polygon), function(navigable, non navigable}, {Connect(Waterbody)}, {geo-entity(bed of watercourse, waterlogged area)}>

Suppose that this user searches a network of geospatial databases for "watercourses" in the context of "dryness." Consider the concept "stream" which is included in the ontology of another database of the network.

C2: <**stream**, {depth(low, high), spatial extent(surface, moving surface), role(navigable, non navigable)}, {Meet(Lake)}, {geo-entity(bed of watercourse, waterlogged area)}>

First, with no views being defined, and therefore no contexts being specified, we are unable to find if "stream" and "watercourse" can be in a similar context of "dryness". With a lexical matching approach, we would however find pairs of synonyms: "watercourse" ↔ "stream", "polygon"↔ "surface", "connect" ↔ "meet", "waterbody" ↔ "lake", "function" ↔ "role". With semantic mapping rules such as those that were presented in (Bakillah et al. 2009), we would find that "watercourse" overlaps "stream," but note that we would be unable to identify that water level corresponds to depth since those properties are not lexically related. Now consider that we employ the MVA generation method we have developed and we build MVACs for "watercourse" and "stream". Suppose we have extracted two views for the concept watercourse, corresponding to contexts dryness, and flooding:

MVAC1: Watercourse
View1(watercourse): Context(**dryness**) → {water level(low), spatial extent(polygon), function(non navigable)}}, {Connect(Waterbody)}, {geo-entity(bed of watercourse)}>

View2(watercourse): Context(**flooding**) → <watercourse, {water level(high), spatial extent(moving polygon), function(navigable)}, {Connect(Waterbody)}, {geo-entity(waterlogged area)}>.

In addition, the following dependencies are extracted for "watercourse:"

{(d1:water level(watercourse, low)→ function(watercourse, not navigable), (d2:water level(watercourse, high)→ function(watercourse, navigable)}

For clarification, this example is shown in Figure 5.3.

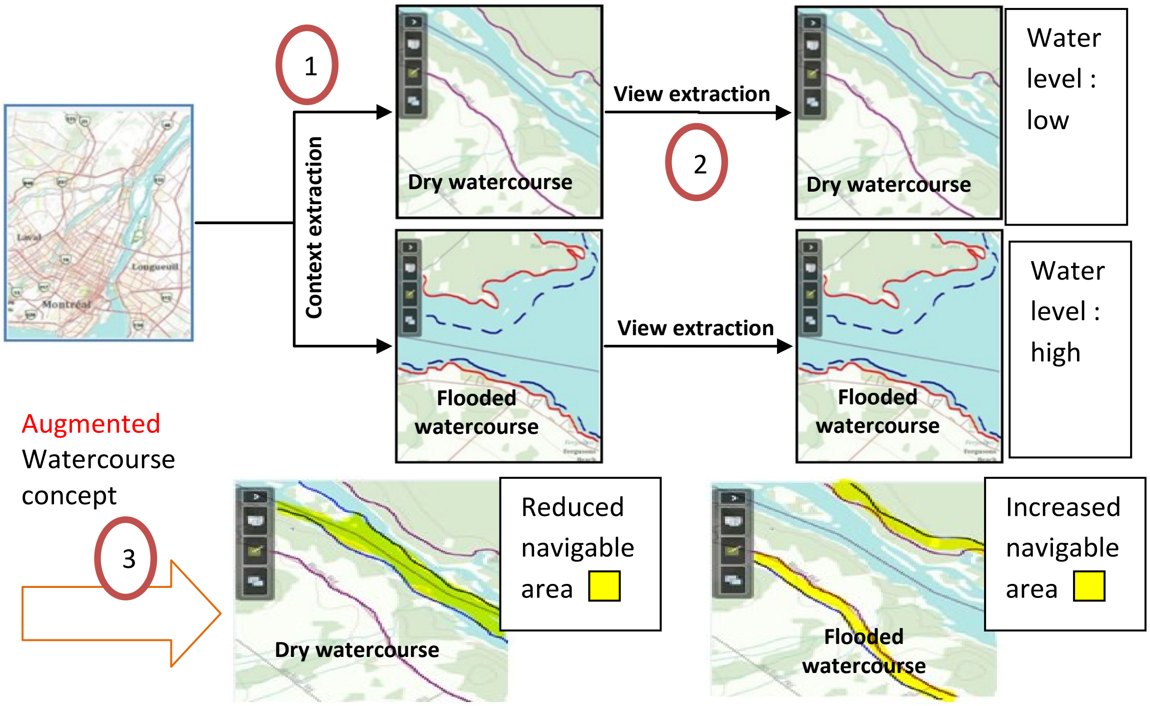


Figure 5.3 Example of semantically augmented concept "watercourse"

For the concept "stream", we have for example extracted:

MVAC2: Stream
View1(stream): Context(**lack of rain**) → <**stream**, {depth(low), spatial extent(surface), role(non navigable)}, {Meet(Lake)}, {geo-entity(bed of watercourse)}>
View2(stream): Context(**rain season**) → <**stream**, {water level(high), spatial extent(moving surface), role(navigable)}, {Meet(Lake)}, {geo-entity(waterlogged area)}>

{d3:(depth(stream, low)→ role(stream, not navigable), (d4:depth(stream, high)→ function(stream, navigable)}

Now we show how the MVAC enables to improve query results by detecting implicit matches, using the structure of the MVACs. After having deduced the lexical matches indicated above, comparing the different dependencies of C1 and C2, we find that d1 has the same structure as d3, and d2 the same structure as d4, which allow proposing the following match: Water level↔Depth. We were able to find this match only because we have augmented the concept with dependencies, which have enriched the concepts' structure. Comparing the contexts of the different views of "watercourse" and "stream" from a lexical-based approach does not allow finding that "lack of rain" corresponds to "dryness." However, if we compare the definitions of **View1**(stream) and **View1**(watercourse), knowing the previous matches, we find that **View1**(stream) is equivalent to **View1**(watercourse), resulting in the following match: Context(**lack of rain**) ↔ Context(**dryness**). This allows the user finally to retrieve "stream" as a concept similar to "watercourse" in the context of dryness. This example shows that augmenting the concept with new structures, i.e., views and dependencies, can help to match concepts, contexts or features of concepts that seem dissimilar, and therefore improve semantic interoperability between geospatial databases.

5.9 Conclusions

In this paper, we argued that to improve semantic interoperability approaches, one main problem that must be tackled is the poor definition of concepts. This is especially true regarding the geospatial domain where concepts are defined by spatial and temporal features, in addition to multiple contexts and implicit dependencies between features. To address this issue, we have proposed the Multi-View Augmented Concept Model (MVAC), and a MVAC generation approach that includes a view extraction and semantic augmentation methods. We have shown that with the MVAC, we can improve semantic interoperability because we can discover more semantic relations between concepts of different ontologies. Therefore, the MVAC can play an important role in a global semantic interoperability approach designed for ad hoc networks where ontologies of databases are very heterogenous, such as in disaster management and in the environmental and health domains. In future work, we will consider the MVAC as a basis for such an approach, with the goal of developing a semantic interoperability approach that is adapted to the MVAC model, since the quality of semantic interoperability depends on the ability of the semantic mapping approach to consider all the characteristics of the input concepts (Bakillah et al. 2008).

CHAPTER 6

Real Time Query Propagation Strategies with Lightweight Coordination Calculus for Ad Hoc Networks of Geospatial Databases

M. BAKILLAH, M.A. MOSTAFAVI

Accepted for publication in the *Journal of Network and Computer Applications*

6.1 Présentation de l'article

Ce chapitre présente le dernier composant du cadre conceptuel, soit l'approche temps réel de propagation de la requête qui a été introduite au Chapitre 3.

Une multitude de sources de données géospatiales sont maintenant rendues disponibles par l'entremise de réseaux dynamiques, tels que les réseaux ad hoc; par conséquent, de nouvelles approches plus adaptées sont requises afin de pouvoir propager les requêtes géospatiales aux sources pertinentes du réseau, tout en prenant en compte les aspects géospatiaux. Bien que plusieurs approches pour la propagation des requêtes existent, elles utilisent différents critères pour sélectionner les sources pertinentes, et plusieurs d'entre elles s'appuient sur les mappings sémantiques existants entre les sources, alors que dans un réseau ad hoc, les sources sont autonomes et peuvent entrer et quitter le réseau dynamiquement. Un de nos objectifs, lors de la propagation de la requête, est plutôt de réduire le nombre de sources auxquelles il faudra accéder pour répondre à une requête aux seules sources pertinentes, ce qui réduira par le fait même la quantité de mappings sémantiques qu'il faudra calculer pour traiter la requête. Dans cet article, nous présentons des stratégies temps réel de propagation des requêtes qui visent à répondre à ces besoins. Ces stratégies reproduisent le comportement des membres d'un réseau social lorsqu'ils communiquent entre eux et qu'ils disséminent de l'information. Ces stratégies s'inscrivent dans un processus d'interopérabilité sémantique en temps réel dans un réseau ad hoc de bases de données géospatiales. Les stratégies sont formalisées avec le Lightweight Coordination Calculus (LCC), qui permet les interactions basées sur des normes sociales et des contraintes dans un système distribué. L'implantation des stratégies et les expérimentations menées démontrent que les stratégies se complètent entre elles afin de fournir une réponse optimale à la requête.

6.2 Presentation of the Article

This chapter presents the last component of the framework, that is, the real time query propagation approach which was outlined in the framework presented in Chapter 3.

Geospatial information is increasingly made accessible in large volume through dynamic networks such as ad hoc networks and consequently adapted approaches are required to propagate geospatial queries to relevant sources, while taking into account geospatial aspects. While different query propagation approaches exist, they each use a different set of knowledge to select relevant sources, and many of them rely on existing semantic mappings between sources, whereas in ad hoc networks, sources are supposed to move autonomously and in a dynamic manner. Our goal for query propagation is rather to reduce the number of sources that must be accessed, and therefore the volume of semantic mappings that must be computed to process the query. In this paper, we propose three real time query propagation strategies to address these problems. These strategies aim to imitate the communication and dissemination behaviours of members of social networks. The strategies are meant to be integrated into the proposed real time semantic interoperability framework for geospatial databases of ad hoc networks. Strategies were formalized via the Lightweight Coordination Calculus (LCC), which supports distributed interactions based on social norms and constraints in networks. The implementation and testing of the strategies show that they complement each other to provide optimal query answer.

6.3 Introduction

In recent years, the volume of geospatial data has increased significantly due to the development of acquisition technologies. At the same time, the development of communication technologies has led to the spreading of different kinds of dynamic network, including ad hoc networks, where sources may enter or leave the network, and groups may form or dissolve in a dynamic and indeterminate fashion. Examples of ad hoc networks include networks formed by mobile, spatially-located devices, and geo-sensor networks that monitor, for instance, environmental phenomena. The accessibility of network technologies has exacerbated the need for tools that can support the dynamic establishment of collaborations among various actors, and often for very specific needs, for instance, following a nation-wide disaster where help must be coordinated between international stakeholders. However, establishing collaborations and enabling geospatial data sharing in this kind of dynamic setting is very challenging. On the one hand, one of the main issue is to locate the sources of the network that are able to answer a given data request. On the other hand, it is very likely that more than one source will be required to answer the request. In addition, a sound decision can only be made when information is corroborated by several sources; consequently, individuals and organisms share data because they want to reduce uncertainty and the risk of misinterpretation (Fox 2008).

Several approaches have been proposed to find relevant sources that fit users' requirement in a network (Staab et al. 2004; Zhuge et al. 2004; Zeinalipour-Yazti *et al.* 2005; Giunchiglia and Zaihrayeu 2006; Mandreoli et al. 2006; Haase et al. 2007; Wiegand and Garcia 2007; Montanelli and Castano 2008). Typical approaches are based on ontologies, which are commonly defined as a "formal specification of a conceptualisation" (Gruber 1993), or simply put, a (more or less) formal representation of concepts and relations that describe a domain of interest (Agarwal 2005). Semantic mappings or semantic similarity values are used to compare concepts contained in the request with concepts of ontologies describing data sources of the network. If concepts are close in meaning, it is inferred that the source in question is appropriated to answer the query. Some of those approaches rely on a central repository of knowledge containing all semantic mappings (Staab *et al.* 2004; Bai *et al.* 2009). In the context of an ad hoc network, this kind of approach is not scalable because of the important volume of sources. In addition, computing all semantic mappings is not an acceptable solution as it is a costly task.

Rather, an approach that aims at finding the most relevant sources while computing a smaller volume of semantic mappings (or, equivalently, requesting the smallest number of sources) is required. The goal of query propagation approaches is to find the relevant sources of a network to which a given query should be forwarded, in order to ensure the optimal query answer. Since networks are significantly large, existing query propagation approaches usually assume that the network is decentralized. This means that there is no central server or "authority" that has a global knowledge on the sources available in the network, and that could offer central repository and communication services; rather, queries are propagated from one source to another (Cudré-Mauroux 2006; Montanelli and Castano 2008). As presented in our literature review in this paper, there are already some approaches that have been developed for query propagation and that were dedicated to decentralized networks. In this paper, we proposed a framework that aligns with those approaches, but whose contribution is to propose three complementary real time strategies for geospatial query propagation in ad hoc networks of geospatial databases. In addition, in the geospatial domain, queries have a spatial and temporal component, i.e. the user is looking for data that pertains to a given geographical location and period in time. This means that during query propagation, in order to select the appropriate query recipient, we must address the issue of heterogeneous spatialities and temporalities. However, existing query propagation approaches did not consider this aspect of queries. Therefore, to be suitable for ad hoc networks of geospatial databases, our approach takes into consideration the spatial and temporal components of the queries. Another difference with existing approaches is that the goal of the proposed query propagation strategies is to determine an optimal order over the existing databases, to identify the most relevant ones, without having to compute semantic mappings between the query and the concepts of all ontologies. Rather, only once the relevant databases are identified, semantic mappings can be computed between the query concept and the concepts of the selected ontologies to retrieve the requested data. Therefore, one of the advantages of our approach is that a smaller number of semantic mappings will be required to find the requested data.

The proposed strategies are meant to be integrated into our real time semantic interoperability framework, which is based on social network principles. As such, this paper will show that the strategies reproduce some of the key abilities used by members of a social network to communicate and disseminate information in an effective manner. The strategies are formalized in the Lightweight Coordination Calculus (LCC), which is a logic framework that allows agents to interact in a social context with social constraints, in a distributed manner (Robertson 2004). A prototype implementation and experimentation shows the performance of individual strategies and suggests that best performance can be achieved through their combination.

6.4 Related Work on Query Propagation in Networks

Propagating a query to relevant databases of a decentralized and dynamic network is a challenging issue, since we have to balance, on the one hand, the quality and relevance of query answers, and on the other hand, the efficiency of the approach. In the geospatial domain, users' queries are usually more complex, since they can have spatial and temporal components, for example, "find the flooded areas close to urban areas in the surrounding of Quebec City," or "find the floods that occurred in Quebec City since 1990." Several approaches have been proposed to address the problem of query propagation since the advent of ad hoc and peer-to-peer networks. Ad hoc networks are based on a computing paradigm that enables the rapid, on-the-fly formation and dissolution of networks with short existence, often for short-term purposes. An ad hoc network is composed of nodes that represent autonomous systems. In this paper, we address the issue of query propagation in ad hoc networks of geospatial databases, i.e. a dynamic network where nodes represent geospatial data-producing or data storage devices, including wireless mobile devices, sensors and geo-sensors, etc. However, many approaches that were developed for similar types of networks (e.g., peer-to-peer networks or distributed database systems) are also relevant to this work. Peer-to-peer networks are networks where participants, called peers, have both capabilities of data and service consumer and provider, and may form groups (called super-peers) based on acquaintances. We review some representative approaches which are compared according to six characteristics (Table 6.1):

- type of network;
- representation of the knowledge about nodes of the network;
- formalization of queries; the more the query is rich and complete, the more the query propagation approach may be accurate. This, however, also have an impact on the cost of the propagation;
- criteria for selecting query recipients;

- whether semantic mappings are already computed or computed at run-time;
- whether there is an update mechanism that reacts when the network is modified (for example, when a node is added to the network).

For comparison purposes, our approach is included in Table 6.1. A first category of query propagation approaches includes the approaches that are based on previous interactions between nodes (peers) of the network. The Intelligent Search Mechanism (ISM) proposed by Zeinalipour-Yazti et al. (2005) is a method for information retrieval in peer-to-peer networks. When a peer receives a query, it retrieves the answers to previous queries in order to select only the peers that are the most likely to give a relevant answer to this new query. A query similarity function is employed to compare past queries to the current query. The query is then forwarded to selected peers only. The authors argue that the ISM method is efficient because the propagation is limited by the number of neighbors, and that it is scalable because it requires no global knowledge. The REMINDIN' (Routing Enabled by Memorizing INformation about Distributed INformation) approach proposed by Staab et al. (2004) [where routing, in this context, is another term used to indicate, like the term "propagation," the forwarding of the query to relevant nodes] is another example of mechanism that uses interactions between peers. The REMINDIN' approach allows a peer to assess the confidence value of another peer based on the number of correct answers given by this peer for a query. The confidence value is updated when the peer has new interactions related to the same topic of interest. The query is propagated to the peers who have the highest confidence value. Note that in this approach, each peer maintains an RDF ontology that describes the semantics of data held by the peer, or another kind of conceptual knowledge related to this peer. One characteristic we note with respect to interaction-based approaches is that their performance depends on previous queries, so they may not perform well at the beginning, or have poor performance when a rarer query is submitted. Consequently, we argue that query propagation approaches based on past interactions should be combined with a complementary strategy that uses another kind of knowledge to determine query recipients.

Another category of query propagation approaches includes the approaches that are based on semantic mappings. In the P2P Semantic Link Networks (P2PSLN) approach developed by Zhuge et al. (2004), each peer has an XML schema describing its knowledge. When a peer enters the network, it identifies the semantic mappings between its schema and the schemas of a set of peers using a set of reasoning rules. When a peer receives a query, it forwards the query to relevant peers, which are selected according to the semantic mappings. In the approach proposed by Mandreoli et al. (2006), semantic mappings between concepts of peer's OWL ontologies are computed based on terminological and structural techniques. Query propagation is based on a semantic Routing Index (i.e., an index for query propagation).This index stores the capability of neighbor peers to answer queries, based on semantic mappings. In the H-Link query routing algorithm, the H-Match semantic similarity model is used to find mappings between ontologies of peers and form a semantic overlay that supports query propagation (Montanelli and Castano 2008). When a peer issues a query, it also associates to that query a number of "credits" which restrain the number of peers to which the query can be forwarded. Aberer et al. (2003) have developed a semantic interoperability approach for large decentralized networks that relies only on pairwise, local interactions between peers, and which introduces the principle of quality of query answers. In this approach, it is assumed that semantic mappings were already computed between concepts of peer's ontologies. When a peer receives a query, it has to decide to which other peers of its neighborhood it will send it, based on syntactic and semantic criteria that measure the quality. In addition, new semantic mappings can be derived with the principle of transitivity; it means that queries can be propagated to peers for which no direct mapping exists, a technique they call *semantic gossiping*. The problem with semantic mapping-based approaches is that since they depend on semantic mappings to determine recipient nodes, they do not necessarily reduce the burden of computing a high volume of useless semantic mappings, since they rely on semantic mappings to determine which node of the network should answer a query. In our approach, we argue that instead of relying on semantic mappings between the query and all the concepts that compose an ontology held by a node, we could rely on a semantic relation between the query and the context of a node; this context would encompass the knowledge formalized by the concept of the ontology with more general context parameters.

Finally, some query propagation approaches are based on groups of peers that share a common interest/topic/domain. In Giunchiglia and Zaihrayeu (2006), the notion of interest group was introduced with the purpose of determining, for a given query, the query scope, that is, the set of nodes a query will be propagated to. This means that a query can be propagated only inside one group. This approach also uses semantic correspondences to propagate queries inside the interest group. In the approach of Haase et al. (2007), a shared ontology is employed to describe the expertise of each peer using a common vocabulary. The selected recipients of a query are determined based on the semantic similarity between the query subject and the expertise of peers. In this approach, the existence of interest groups is implicit. In the GNutella system[4] super-peers (groups of peers with a leader peer) are responsible of distributing queries to appropriate subsets of peers (Kornfilt and Hauswirth 2006). When a peer enters the network, it registers to a super-peer and provides its metadata which describes the most significant features of this peer. One of the limitations of propagation approaches based on groups of peers is that they tend to restrain the propagation to a predetermined set of peers. Therefore, we argue that propagation strategies based on groups should be used in combination with propagation strategies that work at the node level.

Table 6.1 Query Propagation Comparative Study

	Cudré-Mauroux 2006	**REMINDIN'**	**P2PSLN**	**ISM 2005**	**Mandreoli et al.**	**H-Link**
Type of network	Peer-to-peer	Peer-to-peer	Peer-to-peer	Peer-to-peer	Peer-to-peer	Peer-to-peer
Knowledge representation	No specific language, concept and properties (RDF, XML)	RDF statements	XML schema	list of the most recent past queries	Concepts as in an ontology, a relational table or XML schema	OWL ontologies with concepts, properties, datatype properties
Queries	Concept and its properties	SeRQL query language	Concept and its properties	Set of keywords	A concept	Concept and properties
Criteria for selecting	syntactic and	Confidence between peers	Semantic mappings	Similarity between	Semantic Mappings	Semantic affinity

[4] Clip2: The Gnutella Protocol Specification v0.4 (Document Revision 1.2) (2001) http://www9.limewire.com/developer/gnutella protocol 0.4.pdf.

query recipients	semantic criteria measuring the quality of mappings	based on previous correct answers given by a peer for a query		current and past queries		based on semantic mappings
Semantic mappings	Already computed	Already computed	Computed at run-time	Computed at run-time	Already computed	Computed at run-time
Presence of update mechanism	no	no	Update of semantic links if arrival or departure of peer	no	no	no

Table 6.1 (continued)

	Giunchiglia and Zaihrayeu 2003	**Haase et al. 2008**	**GNutella 2004**	**Our Query propagation approach**
Type of network	Peer-to-Peer	Peer-to-Peer	Peer-to-Peer	Peer-to-Peer ad hoc network of geospatial databases
Knowledge representation	Concepts and attributes of a database schema	Set of keywords describing source expertise	RDF schemas	Database's context, coalition's context and memory (past queries)
Representation of queries	SQL queries	Set of keywords	RDF statements	Concept and context of query (OWL statement)
Criteria for selecting query recipients	Interest groups	Semantic similarity between query subject and peers expertise	Super-peers	Affinity between contexts of databases, coalitions, or past queries
Semantic mappings	Already computed (semantic correspondences)	Computed at run-time	Already computed	Don't need to be computed between concepts of ontologies
Presence of update mechanism	no	no	Update indices stored by super-peers	Update of propagation graph when peer is added or removed, coalitions changed

In Table 6.1, we note that existing query propagation approaches rely on a single type of knowledge representation and single criterion for selecting query recipients; our contribution is to suggest a framework where three different but complementary query propagation strategies can be deployed. The strategies rely of different kinds of knowledge representation and therefore, they are expected to improve the efficiency and adaptability of the propagation process. In addition, few approaches include an update mechanism; in fact, the existing mechanisms are updating the knowledge on new nodes that enter or leave the network, but not the propagation graph itself. This means that the results of a query propagation process are always final and are not updated when a new source, for example, joins the network soon after the processing of the query. Consequently, existing query propagation approaches are not adapted to continuous queries, i.e. queries specified to run for a given time period and reactive to push-based data streams (Zafeiropoulos et al. 2009). In the proposed framework, we address the issue of network dynamicity by extending the propagation approach with an algorithm for updating the propagation graph.

6.5 Proposed Framework for Real Time Query Propagation in Ad Hoc Networks of Geospatial Databases

The aim of query propagation is to determine the propagation path along which a geospatial query should be forwarded, in order to reach the most relevant databases that can answer this query. The idea behind the proposed framework is to develop query propagation strategies that reproduce the abilities and behaviours used by members of social networks to reach other members that can fulfill their information needs. All query propagation strategies rely on semantics to determine the most relevant databases, but the different strategies use different types of knowledge and can be deployed at different levels. We consider three ways that are used by people in social network to search for the persons that can fulfill their information needs. Firstly, they can look for organizations of people (companies, government department, associations, etc.) whose context corresponds to their information needs. Secondly, at the individual level, they can compare the context of another person with their information need to assess whether the person holds the requested knowledge (for instance, obtain information on the person's job, hobbies, place of living, age, etc). Thirdly, they can obtain information on the kind of information this person has already provided to others (ex: services he offers), to assess whether the person has relevant experience. The three query propagation strategies that are proposed and that correspond to those scenarios are the following:

- The *coalition-based strategy* uses coalitions of geospatial databases that were formed in the network to identify a set of databases agents that can answer a query. A coalition is a set of databases (and their agents) that have common contextual characteristics (such as common domain or function). Relevant coalitions are determined based on a comparison of the context of the query and the context of the coalition.

- The *context-based strategy* identifies relevant databases based on context affinity between query and databases; because it relies on comparison between query and individual databases, it is intended to be used inside the scope of a single coalition that was previously selected.

- The *memory-based strategy* uses the knowledge that databases agents have about queries that were successfully answered to forward queries to other relevant databases agents. It is also developed to be deployed inside a selected, relevant coalition of databases.

Figure 6.1 illustrates how the propagation strategies are deployed in an ad hoc network when a query is issued by a requestor node. For convenience, the term "node" is used to refer to an agent and its database, in addition to the knowledge held by this agent. The ad hoc network is partitioned into coalitions of geospatial databases. Each coalition has a central database agent which is the "gateway" of the coalition. How coalitions of geospatial databases are formed and central database agents are designated is detailed in Chapter 4. The coalition-based strategy is deployed to resolve the problem of *inter-coalition propagation*, that is, how a query can be forwarded from coalition to coalition. When a coalition is selected as a query recipient, the query is sent to the central database agent. Then, we need to resolve the problem of intra-coalition propagation; the context-based or memory-based strategies are used for this purpose, since the coalition-based strategy, as defined above, is designed to work only at the coalition level. The result of the query propagation process is a query propagation graph that determines a partial order over selected nodes that can answer the query. The first nodes (next to the requestor node) in the propagation graph hold the database estimated to be the more relevant to answer the query, while the leaf nodes hold the less relevant ones.

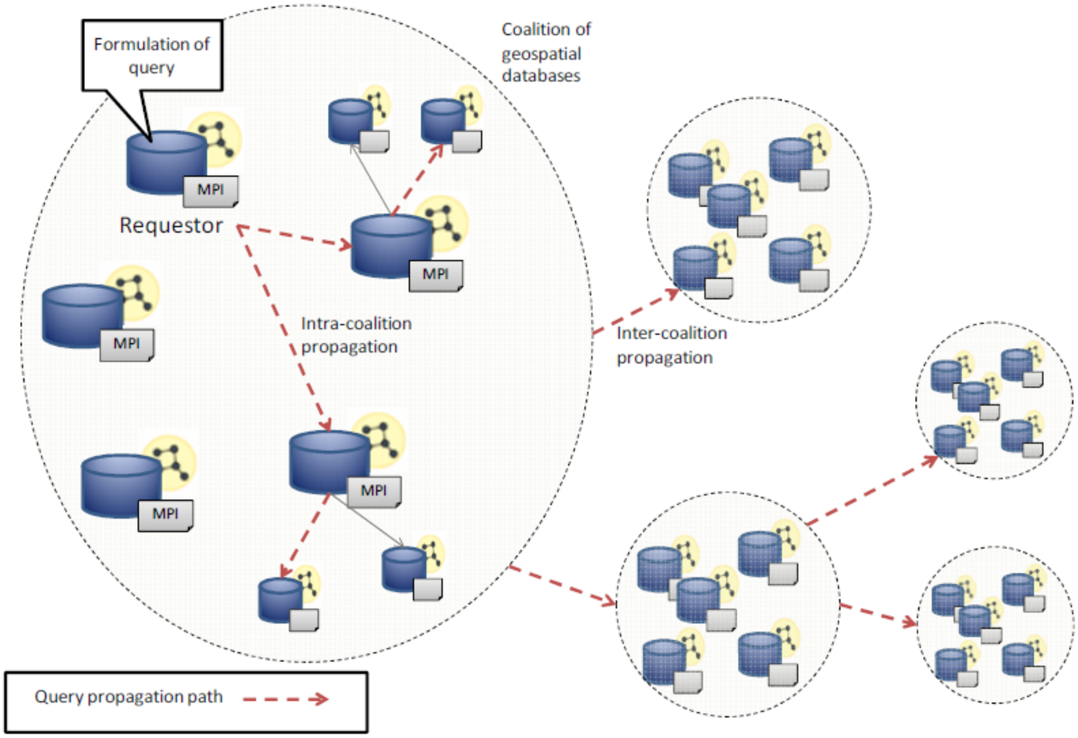

Formulation of query
Coalition of geospatial databases
MPI
Requestor
Intra-coalition propagation
Inter-coalition propagation
Query propagation path

Figure 6.1 Illustration of the intra and inter-coalition propagation

In this framework, we consider real time as referring to the changes that occur in the network, due to the adding or removal of sources from the network, in addition to the formation and dissolution of coalitions. The three strategies are real time strategies because we provide an algorithm that determines how a propagation graph is updated when such changes affect the ad hoc network.

Figure 6.2 illustrates the architecture of the proposed framework, which is composed of three main modules. The Knowledge Representation Module contains and manages the knowledge that can be used to select recipient nodes. Each node holds an ontological description of the context of the database. The context is constituted of characteristics related to theme, space, and time. If the node is a central database agent, it also holds the ontological description of the context of the coalition it belongs to, which is defined by analogous characteristics. The *memory* stores the information about previous interactions. More concretely, this memory is formed by all past queries that has been successfully answered by the node. The Memory Update Module updates the memory when a query is successfully answered. The Query Propagation Module is composed of the three modules responsible for deploying the three strategies, and a strategy manager that receives an entering query and coordinates the strategies accordingly. The Real Time Strategy Adaptation Module determines how a propagation graph can be modified when a change occurs in the ad hoc network. Sending and receiving messages between nodes is ensured by the Communication Module.

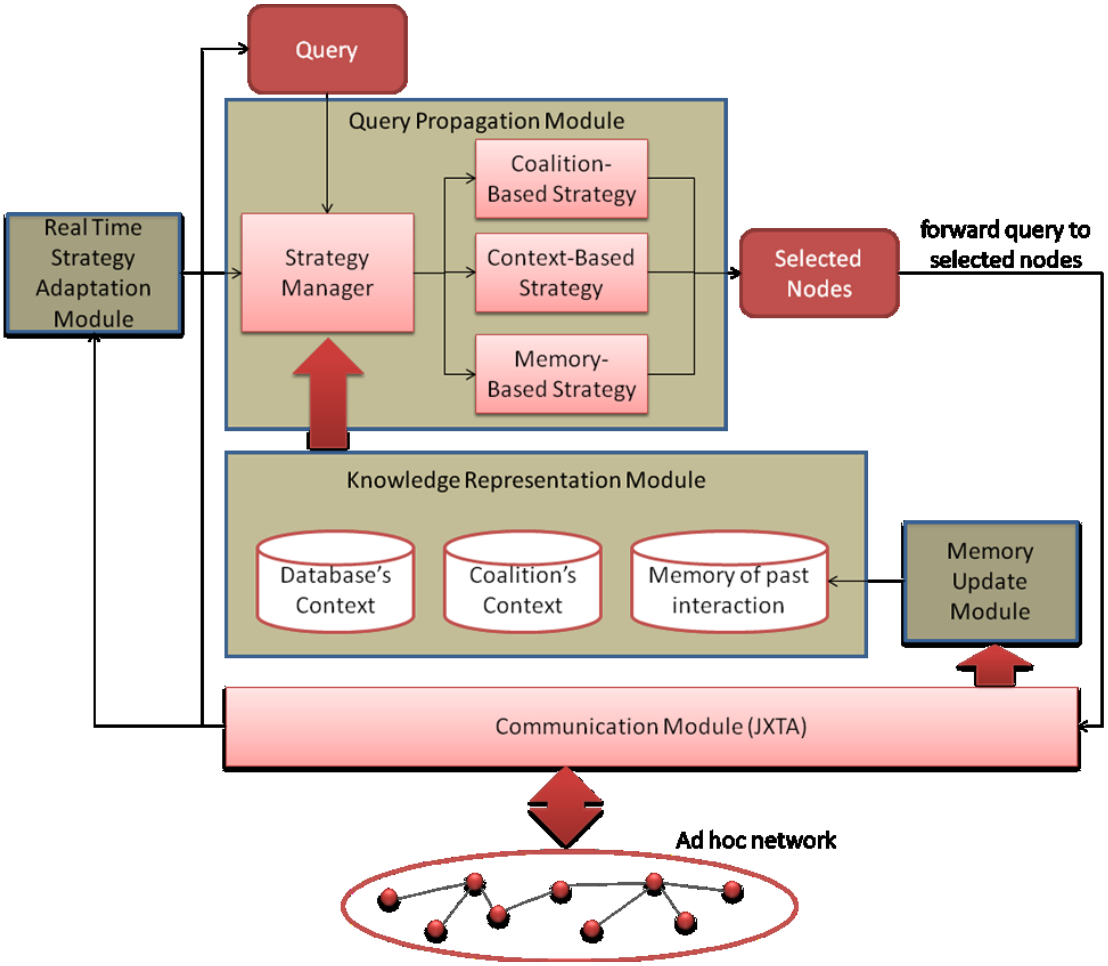


Figure 6.2 Architecture of the proposed framework

The strategies are formalized with the Lightweight Coordination Calculus (LCC) (Robertson 2004). LCC is a suitable framework for formalizing strategies since it is meant to express interactions among distributed processes, such as agents in an ad hoc network. LCC represents interactions as message passing between agents having specific roles in a social context, where the agents' behaviors are determined by conditions expressing social norms. More details on LCC are provided in Chapter 2. LCC is easily implemented with an object-oriented language. In a generic manner, a social norm is defined by an antecedent and a consequent (the predicates that must realize if the antecedent is true). Table 6.2 represents the LCC syntax (Robertson 2004). The => symbol indicates message passing between agents; more specifically, $M \Rightarrow Agent$ means that a message M is sent to the agent, while $M \Leftarrow Agent$ means that a message M from the agent is received. The symbol ← expresses the logic implication. Logic implication dominates message passing operators (=>). *Null* indicates that there is no message passing. The symbol ∧ is the conjunction operator ("and"), while ∨ is the disjunction operator ("or"). Complex agent definitions can be expressed by using the sequential ("then"), choice ("or)", or parallel composition ("par").

Table 6.2 LCC syntax

Framework	$: = \{Clause, \ldots\}$
Clause	$: = Agent :: Adef$
Agent	$: = a(Role,\ Id)$
ADef	$: = null \leftarrow C \mid Agent \leftarrow C \mid Message \leftarrow C$
	$ADef\ then\ ADef \mid ADef\ or\ ADef \mid ADef\ par\ ADef$
Message	$: = M \Rightarrow Agent \mid M \Leftarrow Agent \mid M \Rightarrow Agent \leftarrow C \mid M \Leftarrow Agent \leftarrow C$
C	$: = Term \mid C \wedge C \mid C \vee C$
Role	$: = Term$
M	$: = Term$

In our context, propagation strategies are formalized as a set of norms that help a user agent at a node of the network to determine the next query recipient. Depending on specific norms, a user agent can do nothing (i.e. stop propagation), forward a message (the query), change its role from query recipient to query sender, etc.

6.5.1 Representation of Query Context

While several approaches consider the query as a keyword or a concept taken from the ontology of the requestor, we consider that more relevant results can be obtained if we take into account the situation surrounding the query, that is, the *context* of the query. There are different representations of the context in the literature, and each has been developed for a particular aim. For instance, to disambiguate the meaning of a term, the context may be represented by the words surrounding this term in a text (Finkelstein et al. 2001). Context may also represent background knowledge; for instance, in information retrieval tasks, it may represent personal data about the user, such as its interest and profile (Lawrence 2000; Firat et al. 2007). In context-aware systems, the spatial context is more often represented by the description of the environment and situation in which a device (such as mobile sensor) is used (Dey 2001; Baldauf et al. 2007). Similarly, in the geospatial domain, context can be defined as "information about the surroundings of events, features, and transactions, [where] surroundings can be taken to mean a geographic area" (Goodchild 2009, p. 18). Kokla and Kavouras indicate that in the geographic domain, context refers to "the setting, environment, domain in which an entity or topic of interest exists or occur" (2001, p. 4). More specifically, they consider that the context includes information on geographic concept categories, their properties, relations, and operations. In our approach, the context of the query will be used to represent the purpose or the intention of the user that submitted the query, so relevant databases can be more easily identified. Consequently, a query Q is defined as:

Q = <Concept(C), Context(Ctx)>,

Where the context Ctx is defined as:

Ctx(Q) = <Domain(D), Function(F), Geographical Location(GL), Time Period(T)>.

The domain is the knowledge area targeted by the query, for example hydrology, ecology, urbanism, etc. The domain of a database should encompass the main categories of geographical entities referred to by the concepts that compose the ontology associated to this database. The function is the reason for submitting the query, for example, a query on concept house may have for function "purchasing a house", a query on concept "river" may have for function "assessing flooding risk", etc. The function of the database corresponds to the task that the user wants to perform with the data. This parameter is fundamental, since, according to Wiegand and Garcia (2007), many searches for geospatial data are based on the intended tasks, for example, in case of response to a disaster or for land use or urban planning. The geographical location is the area targeted by the query, for examples houses in Quebec. This geographic area can be represented by a *place*, which is an "area of the

Earth's surface that possesses some form of identity," and that can be included in a gazetteer (Goodchild 2009, p. 18). The time period is the time period targeted by the query, such as "floodings in Canada between 1990 and 2010."

6.5.2 Coalition-based Strategy

As indicated in Figure 6.1, the ad hoc network is partitioned with coalitions of geospatial databases. Each coalition has a context. The coalition context, as defined in Chapter 4, is analogous to the query context. It represents the background information considered as relevant to indicate the knowledge area covered by the coalition. The coalition context includes the domain D(C), the function F(C), the geographical location GL(C) and the time period T(C) for the coalition:

Ctx(C) = <D(C), F(C), GL(C), T(C)>.

Since the ad hoc network is dynamic, new databases can enter or leave a coalition; when such change happens, the coalition's context is recomputed with the operators provided in Chapter 4. The principle of the coalition-based strategy is to use existing coalitions to propagate a query to set of databases that can answer the query, based on a comparison of the context of the query with the context of the existing coalitions. This strategy reproduces the ability of members of social networks to use the structure of society to find people that can fulfill their information needs. We have developed a conceptual model based on the Unified Modeling Language (UML); the model formalizes the elements that participate in the coalition-based strategy; it is illustrated in Figure 6.3.

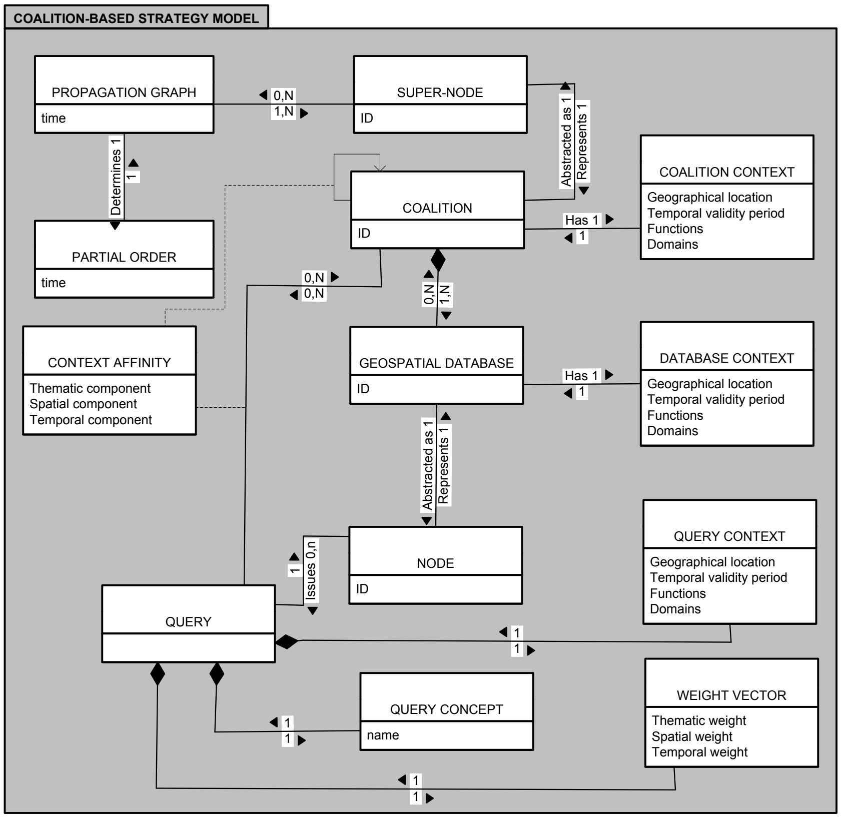


Figure 6.3 Conceptual model of the coalition-based propagation strategy

The propagation graph is formed by super-nodes representing coalitions. Super-nodes are related to each other with labeled arcs, wich indicate:

- The affinity between the contexts of coalitions at related super-nodes, or
- The affinity between the context of the query and a coalition's context at a super-node.

This distinction is fundamental, since it means that the first nodes to which the query is propagated are selected on the basis that they display the highest context affinity with the *context of the query*. However, the next nodes are selected based on their context affinity with the *coalition's context* of previous, intermediary nodes from which they receive the query. The advantage is that nodes whose coalition's context has only indirect affinity with query's context are selected. Therefore, the recall of the strategy, which is the proportion of relevant nodes that were identified (Do et al. 2003), is improved.

We define the context affinity between two coalitions as a weighted sum of affinities between the different parameters of the contexts (domains, functions, geographical location and time period):

$$affinity(ctx1, ctx2) = \frac{w_{domain}}{(1+\delta_{domain}(D(C1), D(C2))} + \frac{w_{function}}{(1+\delta_{function}(R(C1), R(C2))} + \frac{w_{gl}}{(1+\delta_{gl}(GL(C1), GL(C2))} + \frac{w_{t}}{(1+\delta_{t}(T(C1), T(C2))}$$

The principle of this formula is that affinity is an inverse function of semantic distance between corresponding context elements. The user-defined weights assign importance to each type of context element, with w_{domain} + w_{function} + w_{gl} +w_{t} = 1. The definition of semantic distance depends on the type of context elements being compared. Note that when one of the contexts being compared is a query context, the distance between domains includes the normalized sum of the distance between the query concept and the domains. In the following, we describe how the semantic distance is measured for the different types of context element(domains, functions, geographical location and time period).

To measure the semantic distance between domains or between functions, we rely on an upper-level ontology that defines a common vocabulary describing domains and functions and that will allow comparing elements of the contexts of different databases on a common ground. For the purpose of the approach, the ontology of domains and functions, called DFO, was built by gathering the various domains and functions that define the databases of the network, and following a bottom-up approach: creating more specialized categories of functions and domains first, and classifying them into broader categories (Aussenac-Gilles 2005). The upper-level ontology of domains and functions organizes knowledge in terms of concepts, taxonomic relations, and semantic annotations: DFO is a tuple of the form (*C*, *R*, *A*), where:

- *C = {c1, c2...}* is a set of concepts describing domains or functions;

- *R= {r1, r2...}* is a set of taxonomic relationships among concepts of *C*. A taxonomic relationship is a binary relation of the form r(c, c') with c being more general than c'.
- *A= {a1, a2...}* is a set of semantic annotations, where an annotation relates an element of the context of a database, denoted *ε* (*role* or *domain*), to a concept *c* ∈ *C*. An annotation *a* ∈ *A* is written as as a binary relation *a*(*ε, c*).

The semantic distance δ(c, c') between two concepts (nodes) c and c' in DFO is the length (number of relations) of the shortest path between those nodes (Rada et al. 1989). The semantic distance between two geographical locations GL and GL' (defined in the different contexts being compared) can be measured in a similar way, by using an ontology of places, or *gazetteer*, instead of the DFO ontology. A gazetteer defines relations between places, their locations, and their types (Goodchild and Hill 2008). It can also define inclusion relations between places (places that are located inside other places). Considering an ontology of places that contain names of regions (toponyms), synonymy and spatial inclusion relations between toponyms (for example, see Fu et al. 2005 or the SWETO-GS ontology of places of Arpinar et al. 2006), the semantic distance between GL and GL' is also defined as the number of inclusion relations that compose the shorthest path between two geographical relations (places) in the place ontology.

In the current implementation of the query propagation strategies, we used GeoNames[5], a geographical dataset that contains over 8 million geographical names and where location names are associated to coordinates, but also to a type of place (building, city, school, etc.). Places in GeoNames are also linked by inclusion relations. These inclusion relations enable finding geographical locations that are spatially located within other geographical locations (e.g., Laval University is included in Quebec City). The GeoNames database and the spatial inclusion relations provide a comprehensive and valuable source of semantic spatial information that supports the resolution of heterogeneities (naming heterogeneities and heterogeneous levels of spatial granularity) between the geographical locations that are

[5] www.geonames.org/

included in the contexts of compared geospatial databases. Figure 6.4 illustrates the linkage between geographical locations specified in geospatial databases' contexts and the places in GeoNames. Every geographical relation is annotated with a reference in GeoNames.

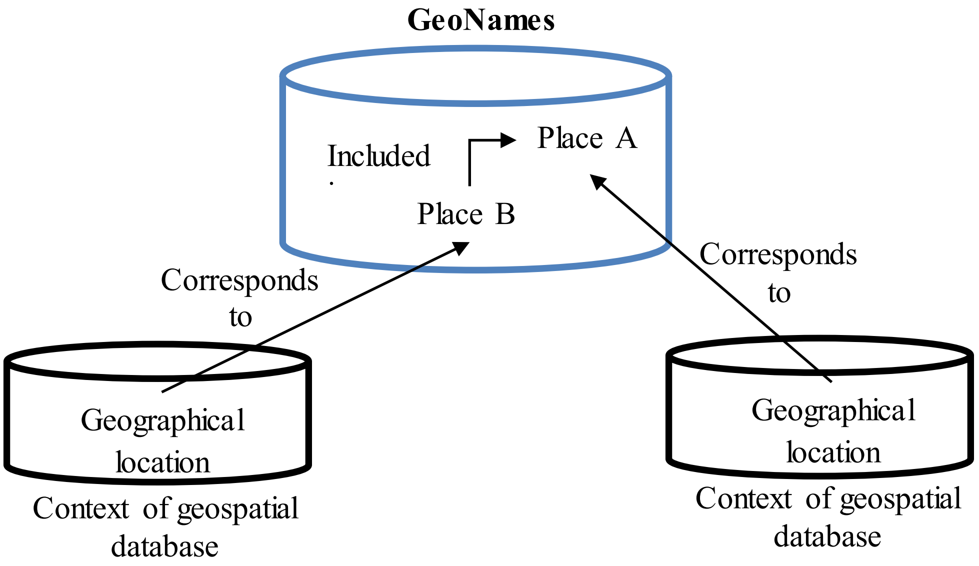


Figure 6.4 Using the GeoNames database to find inclusion relations between geographical locations

Finally, the semantic distance between two temporal periods T and T' (defined in the different contexts being compared), it is measured with the Minkowski distance (for a more complete definition of this distance, see Schwering 2006, p. 65-66). The general expression of Minkowski distance between two elements i and j situated in a space with n dimensions is given by:

$$d_{ij} = \left[\sum_{k=1}^{n} \left|x_{ik} - x_{jk}\right|^{r}\right]^{1/r},$$

where x_{ik} is the value of dimension k for element i and x_{jk} is the value of dimension k for element j. r=1 is the city-block distance and r=2 is the Euclidean distance. We choose r=2 to avoid negative values. Applying this distance to measure the distance between time interval (n=2), we have:

$$\delta_t\big(T(C1),T(C2)\big) = [|t_{11} - t_{12}|^2 + |t_{21} - t_{22}|^2]^{1/2},$$

where t_{11}, t_{12} are the start and end of time period T(C1) of coalition C1, and t_{21}, t_{22} are the start and end of time period T(C2) of coalition C2. Figure 6.5 shows the LCC framework for the coalition-cased strategy. This framework formally describes the interactions that can take place during inter-coalitions propagations.

```
a(requestor(Q), RA) ::
                requestCoalitionContext() => a(coalition_agent(C), A) ←
                        a(networkMember, C) then
                returnCoalitionContext(G(Ctx(C)), T(Ctx(C)), D(Ctx(C)),
                F(Ctx(C))) <= a(coalition_agent(C), A) then
                a(recipient(Q), C) ←localConfidenceInterval(I) ∧
                insideI(ctx_affinity(Q, C)) ∧ maxCycle(False) ∧
                reachThreshold(ctx_affinity(Q, C), Th) then
                Q => a(recipient(Q), C) then a(requestor(Q), A)

a(coalition_agent(C), A) ::
                requestCoalitionContext() <= a(requestor(Q), RA) then
                returnCoalitionContext => a(requestor(Q), RA) ← ¬
                (recipient(Q), C) ∧
                get(G(Ctx(C)), T(Ctx(C)), D(Ctx(C)), F(Ctx(C))) or
                a(recipient(Q), C)) ← Q <= a(requestor(Q), RA) then
                a(requestor(Q), A) ← a(recipient(Q), C)
```

Figure 6.5 LCC framework for the coalition-based strategy

The possible roles of the agents of databases at nodes are *requestor* (the one who formulate a query *Q*), and *coalition_agent(C)* (a central agent that is responsible to manage the coalition *C*). For example, *a(requestor(Q), RA)* means that the agent with id *RA* is the requestor who wants to send the query *Q*. The possible role of coalitions at super-nodes are *recipient(Q)* (by extension, it means that all members of the coalition are recipients(*Q*)) and *networkMember*. At any time, this framework allows the members and coalitions to dynamically change their role.

In the scenario, the requestor sends a *requestCoalitionContext*() message to other coalitions who then send the elements of their coalition context to the requestor. For example, the following sequence: *requestCoalitionContext() => a(coalition_agent(C), A)* indicates that a request to get the context of a coalition *C* was sent to *A*, the coalition agent of *C*. A coalition will be selected as a recipient(*Q*) if:

- for a local confidence interval of context affinity *I*, the context affinity between the query and the coalition (or between two coalitions) is inside the confidence interval *I*. This is expressed by the following clause: *a(recipient(Q), C)* ← *localConfidenceInterval(I)* Λ *insideI(ctx_affinity(Q, C))*. The local confidence interval is computed at each node; it is the interval of context affinity that contains *x* percent of higher values of context affinity between the node and its neigbhour nodes, where *x* can be user-defined. The smaller *x* is, the more selective the algorithm is.
- the maximal number of propagation cycles is not reached (*maxCycle(False)*);
- the context affinity between the query and the coalition (or between two coalitions) is equal or higher than the context affinity threshold (selected by the user). This is expressed by the following clause: *reachThreshold(ctx_affinity(Q, C), Th)*.

When an coalition's role is changed to that of a recipient, the query Q is sent to him, and it is forward to all members of the coalition; the managing coalition agent become a *requestor*(Q) that will send the query to other coalitions.

The advantage of this strategy is its scalability, i.e., its ability to function well even in a large network: to find the relevant nodes to which the query will be propagated, it is not needed to compare the query against each single node of the network, but only against the coalitions' contexts. However, it is less precise than other strategies because the query is sent to every node (databases) inside a coalition, even if some of them may not be relevant. This strategy will display better results if the network is partitioned into several specialized coalitions. The combination of the coalition-based strategy with intra-coalition strategies produces a mixed strategy that can be scalable and precise.

6.5.3 Context-based Strategy

The principle of the context-based strategy is that the query propagation graph is determined according to the affinity between context of the query and context of geospatial databases. It is intended to be used within the scope of a single coalition. It is analogous to the coalition-based strategy; however, it works at the database level. This strategy reproduces the ability of members of a social network to use content at nodes to find people that are able to satisfy their information requirements. In many query propagation approaches, the query is a concept. In the context-based strategy, the context of a query is intended to enrich the query with additional element that helps to determine the spatial, temporal and thematic scope and purpose of the query. Also, the context of a database is intended to help evaluate what kind of queries can be answered, that is, whether data is reusable in a given situation. We have developed a Unified Modeling Language (UML) conceptual model of the context-based strategy, which is illustrated oi Figure 6.6. In this model, each node is associated to a database.

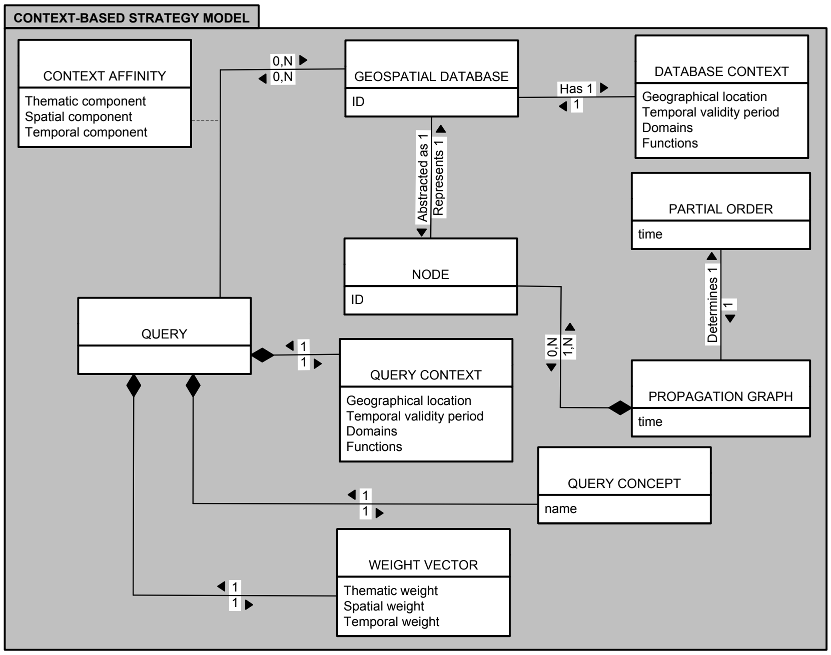


Figure 6.6 Conceptual model of the context-based propagation strategy

The context of a database is formalized as follows:

Ctx(DB) = <D(DB), F(DB), GL(DB), T(DB)>,

where the variables are respectively the domain, function, geographical location and temporal period of the database. In this strategy, nodes are related to each other with labeled arcs, which indicate:

- The affinity between the contexts of databases at related nodes, or
- The affinity between the context of the query issued by a first node and the context of a database at a second node, when the second node is a direct neighbor of the requestor node.

The context affinity between databases is analogous to context affinity between coalitions:

$$\begin{aligned} &affinity(ctx1, ctx2) \\ &\quad = \frac{w_{domain}}{(1+\delta_{domain}(D(C1), D(C2))} + \frac{w_{function}}{(1+\delta_{function}(F(C1), R(C2))} \\ &\quad + \frac{w_{gl}}{(1+\delta_{gl}(GL(C1), GL(C2))} + \frac{w_t}{(1+\delta_t(T(C1), T(C2))} \end{aligned}$$

Figure 6.7 shows the LCC framework for the context-based strategy. This framework formally describes the interactions that can take place during intra-coalitions propagation based on context affinity.

```
a(requestor(Q), RA) ::
        requestDBContext() => a(db_agent, A)←a(coalitionMember, A)
                then returnDBContext(G(Ctx(DB)), T(Ctx(DB)), D(Ctx(DB)),
                        R(Ctx(DB))) <= a(db_agent, A) then
                        a(recipient(Q), A) ←localConfidenceInterval(I)∧
                                insideI(ctx_affinity(Q, DB))∧ maxCycle(False) ∧
                                        reachThreshold(ctx_affinity(Q, DB), Th) then
                        Q => a(recipient, A) then a(requestor(Q), A)
```

```
a(db_agent, A) ::
                    requestDBContext() <= a(requestor(Q), RA) then
                    returnDBContext => a(requestor(Q), RA) ← ¬ (recipient(Q),
A)                  ∧       get(G(Ctx(DB)),    T(Ctx(DB)),    D(Ctx(DB)),
R(Ctx(DB)))
                    or a(recipient(Q), A)) ← Q <= a(requestor(Q), RA)
                    then a(requestor(Q), A) ← a(recipient(Q), A)
```

Figure 6.7 LCC framework for the context-based strategy

The possible roles of the agents of databases at nodes are *requestor* (the one who formulate a query Q), *db_agent* (any passive member of the network that may receive a query), *recipient(Q)* (agent chosen to respond to a query Q), and *coalitionMember* (an agent who is member of the coalition). At any time, this framework allows the member to change their role in a dynamic fashion. In the scenario, the requestor sends a *requestDBContext*() message to members of its coalition who send the element of their database's context to the requestor. A member of the coalition will be selected as a *recipient*(*Q*) if:

- for a local confidence interval of context affinity *I*, the context affinity between the query and the database (or between two databases) is inside the confidence interval *I*;
- the maximal number of propagation cycles is not reached (*maxCycle(False)*);
- the context affinity between the query and the database (or between two databases) is equal to or higher than the context affinity threshold.

When an agent's role is changed to that of a recipient, the query Q is sent to him, and it become a *requestor*(*Q*).

The disadvantage of this strategy is that it is working within the scope of a reasonably small coalition. Otherwise, it is not scalable enough to be applied to the entire network. This is why the coalition-based strategy that works at a higher level of representation of the network is required. However, context-based and coalition-based strategies are both based on static content at nodes. They do not take into account the knowledge that is generated as interactions occur between agents of databases at nodes. This ability is fulfilled by the collective memory-based strategy

6.5.4 Collective Memory-based Strategy

The collective memory-based strategy uses the knowledge that users of databases have about queries that were previously answered to forward queries to the relevant databases. In social networks, people usually keep in their memory the information to find the right people to contact in a given situation. This strategy reproduces the ability of members of a social network to use the knowledge of people to find requested information. We have developed a Unified Modeling Language (UML) conceptual model of the collective memory-based strategy, which is illustrated oi Figure 6.8.

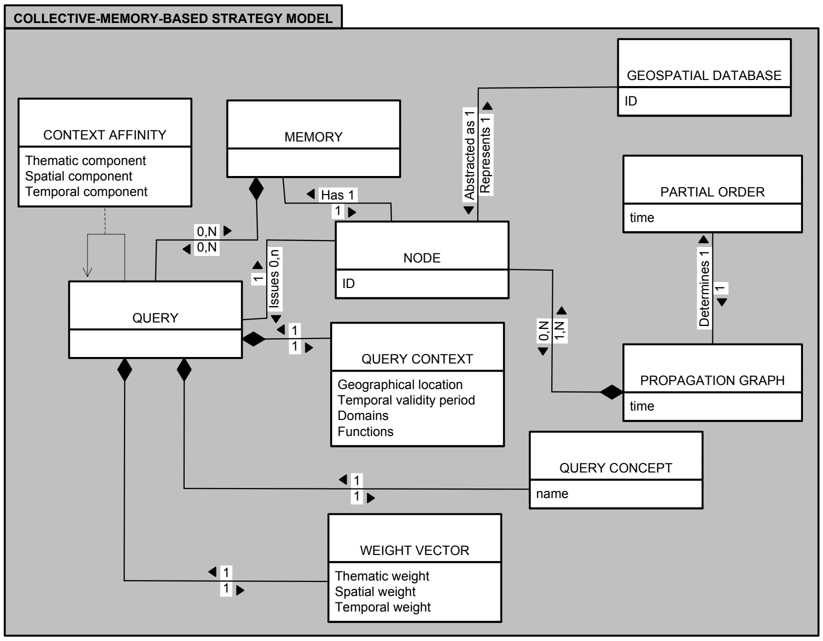


Figure 6.8 Conceptual model of the collective-memory-based propagation strategy

In this strategy, every node (database) is associated to a memory. This memory represents the knowledge about past queries that were correctly answered by this node. Therefore, it is a repository of a set of queries (we note that correctly answered queries can be identified through the feedback between users). Instead of determining the propagation based on context affinity between databases, it is based on context affinity between the current query and the past queries. When a node possesses in its memory a query that is similar to the current query, it can be selected as a recipient node. Figure 6.9 shows the LCC framework for the collective memory-based strategy.

The possible roles of agents of databases at nodes are *requestor*, *db_agent*, *recipient(Q)*, and *coalitionMember*. For example, *a(recipient(Q), A)* means that the agent with id *A* was selected as a recipient of query *Q*. At any time, this framework allows the members to change their role. In the scenario, the requestor sends a *requestLocalMemory*() message to members of its coalition to obtain the set of queries that compose the local memory of each node. For example, the following sequence: *requestLocalMemory() => a(db_agent, A)* indicates that a request to get the local memory of a database agent *A* was sent to *A*. A db_agent will send its local memory only if he has not answered Q previously.

```
a(requestor(Q), RA) ::
        requestLocalMemory() => a(db_agent, A) ← a(coalitionMember,
                A) then returnLocalMemory() <= a(db_agent, A) then
                        a(recipient(Q), A) ←localConfidenceInterval(I) ∧
                        insideI(ctx_affinity(Q, Q(LocalMemory))∧ maxCycle(False) ∧
                                reachThreshold(ctx_affinity(Q, Q(LocalMemory)), Th)
then
                        Q => a(recipient, A) then
                        a(requestor(Q), A)


a(db_agent, A) ::
                        requestLocalMemory() <= a(requestor(Q), RA) then
                        returnLocalMemory() => a(requestor(Q), RA) ←
                        ¬ (recipient(Q), A) ∧ get(LocalMemory) or
                        a(recipient(Q), A)) ← Q <= a(requestor(Q), RA)
                        then a(requestor(Q), A) ← a(recipient(Q), A)
```

Figure 6.9 LCC framework for the collective memory-based strategy

The collective memory is the union of local memories. A member of the coalition will be selected as a recipient(Q) if:

- for a local confidence interval of context affinity *I*, the context affinity between the query and one or more queries of the local memory is inside the confidence interval *I*; this is expressed by the following clause: *a(recipient(Q), A) ← localConfidenceInterval(I) ∧ insideI(ctx_affinity(Q, Q(LocalMemory))).*
- the maximal number of propagation cycles is not reached (*maxCycle(False)*);
- the context affinity between the query and one or more queries of the local memory (or between two queries of different local memories) is equal or higher than the context affinity threshold. This is expressed by the following clause: *reachThreshold(ctx_affinity(Q, Q(LocalMemory)), Th).*

When an agent's role is changed to that of a recipient, the query Q is sent to him, and he becomes a *requestor*(Q). Similarly to the context-based strategy, this strategy works at the database level, so it displays the same disadvantages regarding scalability. However, it is also complementary to the two other strategies because it is based on another kind of available knowledge. Note also that this approach will perform better with time, that is, as more queries are submitted to the ad hoc network. To better illustrate how propagation graphs are computed using any of the three strategies, in Figure 6.10, we show the generic algorithm of the propagation strategies. The algorithm is generic because it is applicable to all three strategies. The process always starts with the formulation of a query by an agent of a database at a node (step 1). The algorithm performs a sequence of cycles, or "jumps", either from node to node, within the scope of a coalition (for context and memory-based strategies), or from super-node to super-node, within the whole network (for coalition-based strategy). "Jumps" or "cycles" refer to the action of forwarding a query from one node to another. Therefore, the following description of the algorithm also applies to super-nodes, replacing nodes with super-nodes. The algorithm is parameterized with a maximum number of cycles, whose role is to avoid the unstopped propagation of the query. If the maximum number of cycles is not reached (step 2), a list of requestor node is created (step 3). A requestor node is a node that will forward the query to other nodes. At the first propagation cycle, the list of requestor nodes contains only the node that initiated the query. At the next propagation cycles, the list of requestor nodes will be filled with the next nodes that were selected as query recipients.

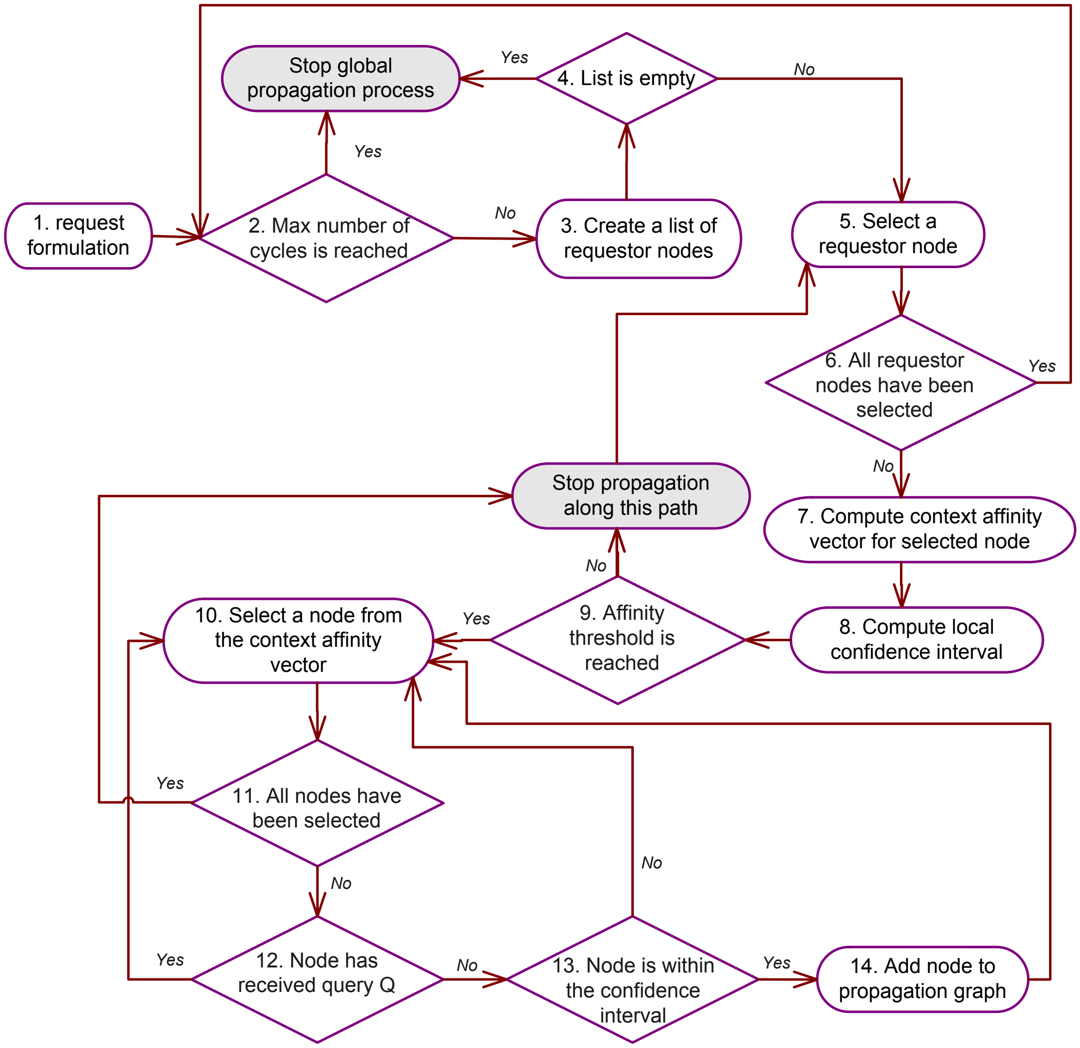

Stop global propagation process
Yes
4. List is empty
No
Yes
1. request formulation
2. Max number of cycles is reached
No
3. Create a list of requestor nodes
5. Select a requestor node
6. All requestor nodes have been selected
Yes
No
Stop propagation along this path
7. Compute context affinity vector for selected node
No
10. Select a node from the context affinity vector
Yes
9. Affinity threshold is reached
8. Compute local confidence interval
Yes
11. All nodes have been selected
No
No
Yes
12. Node has received query Q
No
13. Node is within the confidence interval
Yes
14. Add node to propagation graph

Figure 6.10 Flowchart of the generic propagation strategy algorithm

For each requestor node, a context affinity vector is computed (step 7):

$$V = (a_1, a_2, \dots a_n)$$

This vector is a list of context affinity values between the query and the context of databases, coalitions or queries stored in the collective memory, depending on the strategy being applied at that time. A local confidence interval is computed, which is an interval containing the highest affinity values for this node (step 8). This interval is said to be local because for every node being reached during the propagation process, a different confidence interval is determined dynamically. Only the next nodes for which the affinity with the current requestor node falls within this confidence value will be selected as query recipients (step 9). In addition to the local confidence interval, a query is further propagated only if a global, user-defined, context-affinity threshold is reached. Therefore, if the local confidence interval is below the threshold, the query is not propagated further along this current requestor node. Only the nodes that meet these criteria are selected as query recipient and added to the propagation graph (step 14). To ensure that no node answers twice the same query, the query is given a unique identifier captured and stored by the nodes who received the query. If a node receives a query he has already received (step 12), it stops the local propagation process (but the propagation may continue along other paths). Consequently, the only global stop criterion is the maximal number of cycles. At the end of a cycle, every node that was selected as a query recipient becomes a requestor node and is added to the list of next requestor nodes that will try to forward the query.

6.5.5 Combination of the Query Propagation Strategies

Table 6.3 highlights the features and advantages of the three strategies and the situations in which they can be used. It also indicates how the three strategies can be used in combination in order to maximize the performance of the query propagation process.

Table 6.3 Comparison of query propagation strategies

Query propagation strategy	Features and Advantages	Usage
Coalition-based strategy	• Ability to reach a large number of nodes without having to access the knowledge on every node • Query may be propagated to irrelevant nodes that are included in a selected coalition	• To find groups of databases that could answer a set of related queries • As a first step of the query propagation process, to identify the groups that may contain the most relevant nodes
Context-based strategy	• Ability to identify the relevance of nodes based on the knowledge on the data they hold • More precise than coalition-based strategy • Less scalable than coalition-based strategy	• To find relevant nodes within a selected coalition • To find relevant nodes in a relatively small network
Collective-memory-based strategy	• Ability to identify the relevance of nodes based on the queries that the node was able to answer • More precise than coalition-based strategy • Less scalable that coalition-based strategy	• To find relevant nodes within a selected coalition • To find relevant nodes in a relatively small network • To use when a sufficient amount and a variety of queries were submitted in the network

As illustrated oi Figure 6.1, the three strategies can be combined as follows: when a query is submitted, it is first propagated to the relevant coalitions using the coalition-based strategy. Then, the query must be propagated to the relevant databases inside the selected coalitions with a combination of the context-based and memory-based strategies. To do so, for each node that received the query and that is currently acting as a requestor node, instead of computing a single affinity vector (step 7 of the propagation algorithm), two affinity vectors are computed:

- one containing context affinity values and
- the other containing the affinities between the current query and the queries stored in the memory.

Then, the nodes that will be selected as recipient nodes are the ones that satisfy the selection criteria either based on the context affinity or the memory affinity.

6.5.6 Real Time Adaptation of Propagation Strategies in Ad Hoc Networks

In this section, we propose an algorithm that updates the propagation graph when changes affect the ad hoc network. Changes in the network happen when new goals are determined and as a result, new coalitions are formed; when the goal of a database agent is modified, etc. These changes may result in the addition of a database to the network or to a coalition, the formation of a new coalition, the division of a coalition into several coalitions, the fusion of several coalitions, and the removal of databases from a coalition (coalition shrinking). This algorithm proposes a method that assesses the expected impact of a change before updating the propagation graph. If the change is expected to have a minimal impact on the propagation graph, the change will be ignored; however, the propagation graph will be updated if the change is expected to change significantly the results of a query that was submitted previously. This algorithm is representative of the ability of social network members to deal with frequent changes that affect their network in order to maintain stability or perform only the changes that likely to have a significant positive impact.

The real time strategy adaptation algorithm is presented in Figure 6.11. In this algorithm, the context affinity may be computed by comparing queries, databases' contexts or coalition's contexts. The algorithm first deals with coalition expansion, that is, when a new database enters the network and is added to a coalition where a query propagation graph is valid. This supposes that query propagation graphs have a time validity period that corresponds to the lifetime of the query. A propagation graph whose validity period is expired does not need to be updated. When a coalition expansion is detected (when the coalition expansion operator is employed) (step 1), for each new database, a context affinity vector is computed between the context of this database and the contexts of databases that are already part of the propagation graph (step 2). The database with maximal context affinity is selected (step 3). The fourth step aims at determining if the impact of the new database is sufficient to be taken into account. The distance between the database selected in step 3 and the starting node of the propagation graph is computed. This distance is the number of nodes that separates two nodes in the propagation graph. Since no nodes can be selected twice in the propagation algorithm, this distance is unique (there is no cycle within the graph). If the distance is higher than a maximal distance threshold (step 5), it is considered that the new database will not have a significant impact on the query answer, and the database is rejected from the propagation graph (step 6). Otherwise, the new database is added as the next neighbor node of the node selected in step 3, and the propagation algorithm is started with the new node as requestor node. Note that the only applicable strategy in the case of the adding of a new database is the context-based strategy, since, when a node enters the network, it has no memory of past queries. As shown at the bottom of Figure 6.11, the same update algorithm is applicable at the coalition level, following the creation of a new coalition, either through coalition formation, division, or merging operators.

```
Real Time Strategy Adaptation Algorithm

   When a new Database is added to the coalition:
   1  IF Detection of new node $DB_{NEW}$ in the coalition
   2  THEN Compute a context affinity vector $V_{NEW}$ between $DB_{NEW}$ and
      nodes of existing propagation graph G
   3  Select node $DB_{MAX}$ of the existing propagation graph G with
      maximal context affinity $A_{MAX}$
   4  IF maximal context affinity $A_{MAX}$ is higher than context affinity
      minimal threshold $A_{MIN}$
   5  THEN Compute distance $\delta$ between first requestor node in G,
      $DB_{REQ}$, and node $DB_{MAX}$
   6  IF distance $\delta$ is higher than distance maximal threshold $\delta_{MAX}$
   7  THEN end
   8  ELSE add new node $DB_{NEW}$ to propagation graph G as next neighbour
      of node $DB_{MAX}$
   9  Start propagation algorithm of selected strategy S with new
      node $DB_{NEW}$ as single requestor node.

      When a new coalition is formed in the ad hoc network:
   1  IF Detection of new coalition $C_{NEW}$ in network
   2  THEN Compute a context affinity vector $V_{NEW}$ between new
      coalition and super-nodes of existing coalition-based strategy
      propagation graph G
   3  Select super-node $C_{MAX}$ of propagation graph G with maximal
      context affinity
   4  IF maximal coalition context affinity $A_{MAX}$ is higher than
      context affinity minimal threshold $A_{MIN}$
   5  THEN Compute distance $\delta$ between first requestor super-node,
      $C_{REQ}$, and super-node $C_{MAX}$
   6  IF the distance $\delta$ is higher than distance maximal threshold $\delta_{MAX}$
   7  THEN end
   8  ELSE add new super-node $C_{NEW}$ to propagation graph G as next
      neighbour of super-node $C_{MAX}$
```

```
Start coalition-based propagation algorithm with new super-node
$C_{NEW}$ as single requestor super-node.
```

Figure 6.11 Real time strategy adaptation algorithm

6.6 Implementation and Simulation in Multi-agents System

The main goal of the simulation is to assess and compare the performance of the different propagation strategies to demonstrate in which context each of them is useful. For this, the strategies were implemented in Java on the JXTA platform, which is an open source Java platform simulating a dynamic, open network of peers that may form groups. We have created a set of a hundred ontological descriptions of databases, starting with specifications of existing geospatial databases from the domain of topography, hydrography, road networks, etc., and have introduced random variations within the descriptions. Each description was associated with a node (peer) of the network. Each node also maintains a file for the storage of previous queries, which were formulated using concepts of specifications of databases. Then, we created small coalitions of nodes. Context affinity values are computed on-the-fly by the requestor node.

More specifically, the goal of the experimentation is to evaluate the ability of the propagation strategies to retrieve the relevant databases that can answer a query. For each query, the set of relevant nodes was manually identified. The ability to retrieve relevant nodes is measured with the recall, which corresponds to the ratio between the number of relevant databases that are part of the propagation graph, and the number of manually determined relevant databases. In terms of recall, it is expected that the coalition-based strategy may perform better. The other goal of the experimentation is to evaluate the semantic accuracy of the propagation strategies, that is, their ability to retrieve relevant databases while discarding irrelevant ones. The semantic accuracy is defined as the ratio of relevant databases that are part of the propagation graph and all the databases that are part of the propagation graph. With respect to semantic accuracy, it is expected that intra-coalition strategies will perform since they work at a lower level of detail than inter-coalition strategy. Figure 6.12a to 6.12c shows the results of recall and accuracy for the three strategies. Recall and accuracy are measured against the number of databases (nodes) being reached at some time during the propagation process. In the three strategies, both measures are somewhat independent from the number of nodes being reached, meaning that their performance is stable. In the context-based strategy, accuracy is generally lower than recall, but not as much as in the coalition-based strategy, as expected. The accuracy of the coalition-based strategy is necessarily lower, since it retrieves all databases of a coalition without consideration for their individual description. But, as shown later, this makes the coalition-based strategy more scalable. Still, the experiment confirms that recall is higher in the coalition-based strategy, but it is counterbalanced by low accuracy situated in the interval 0,40-0,57.

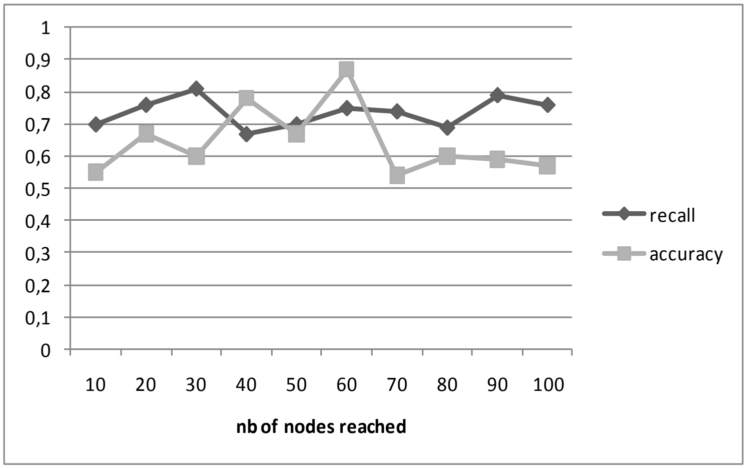


Figure 6.12a Recall and accuracy of the context-based strategy

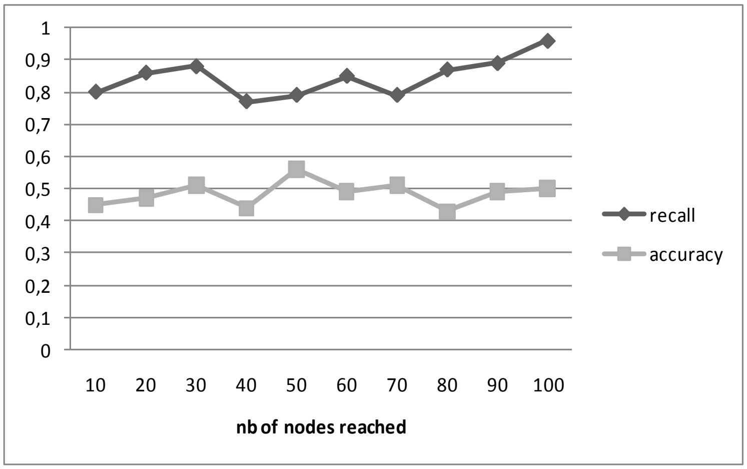


Figure 6.12b Recall and accuracy of the coalition-based strategy

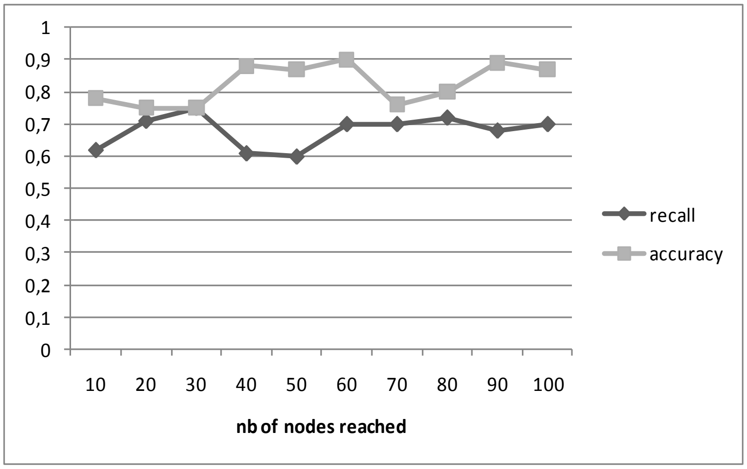


Figure 6.12c Recall and accuracy of the collective-memory-based strategy

The memory-based strategy, which was tested after a learning phase where several, random queries were submitted in the network, offers a better performance in terms of accuracy, with values situated between 0,73 and 0,90. We expect that this is due to the fact that when this strategy is applied, databases that are selected as query recipients are only the ones whose memory contains very similar queries to the current query. Therefore, the accuracy of this strategy will depend on the context-affinity threshold chosen: a lower threshold is more likely to increase recall but reduce accuracy. Those results suggest that the three strategies are complementary to offer better performance. The coalition-based strategy can be used to identify the first coalitions to which the query will be sent; then, a combination of intra-coalition strategies can be used to define the propagation graph inside the selected coalitions. The result is a two-level propagation graph.

The strategies were also again tested against the maximal number of cycles. The goal of this second experimentation is to assess the scalability of each strategy. The total recall is measured at each propagation cycle, that is, the number of relevant databases that were added to the propagation graph with respect to the total number of relevant databases. Therefore, as the maximal number of cycles is increased at each simulation, the recall also increases but not necessarily the semantic accuracy. Figures 6.13 and 6.14 show the recall and semantic accuracy versus the maximal number of cycles for each strategy.

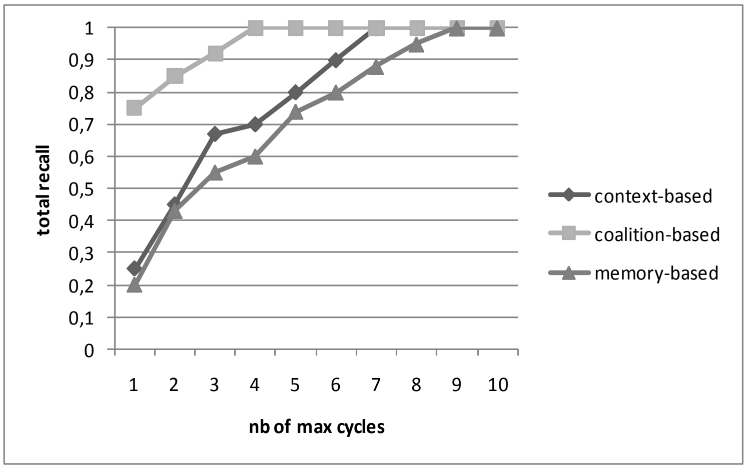


Figure 6.13 Total recall with respect to maximal number of cycles, for different strategies

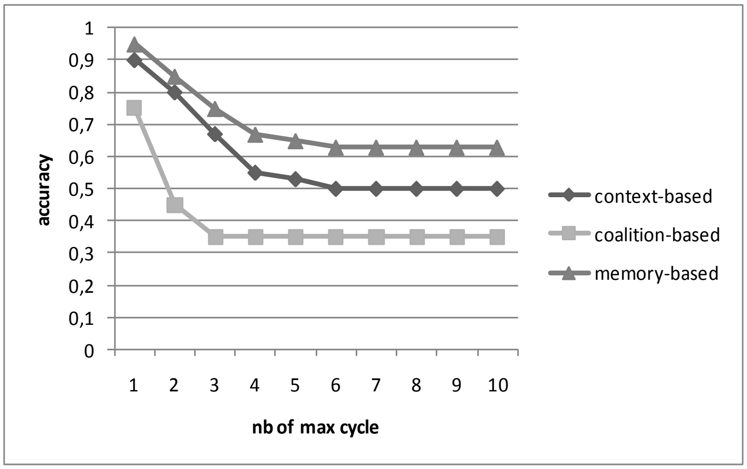


Figure 6.14 Accuracy with respect to maximal number of cycles, for different strategies

We observe that eventually, as the maximal number of cycles is increased, the recall reaches a maximal stable value near 1, while the semantic accuracy is decreasing to reach a lower minimum. The other expected observation is that recall rapidly reaches a maximum when the coalition-based strategy is applied. The stable minimum indicates that even if the maximal number of cycles is manually increased, the propagation algorithm is stopped at some nodes by other parameters, including minimal context affinity. However, the maximal recall of 1 is only reached at around 8 and 9 cycles for intra-coalition strategies, while the accuracy has already reached a minimum. This shows that the propagation strategy is still efficient in the last cycles of propagation. This result again depends on the minimal affinity threshold value which must be high enough to reproduce this behaviour. In further experiments, we plan to assess the behaviour with respect to different thresholds. The overall results of the experimentations show that the strategies are complementary in achieving best performance and that one's weaknesses are counterbalanced by the other's strength. This is representative of social network communication where different strategies are used in relevant contexts to maximize one's knowledge.

6.7 Conclusion and Future Work

The advent of ad hoc networks has increased the access to large volumes of geospatial data for users and decision-makers, whose number is also constantly growing. In turn, this has introduced new challenges with respect to the efficiency of the source discovery process. In this paper, we have dealt with some problems related to propagating geospatial queries to relevant geospatial databases in ad hoc networks. The main goal was to develop query propagation strategies adapted to ad hoc networks that use different social network communication abilities. Our contribution is to propose three strategies that work at different levels, using complementary information in order to maximize flexibility, scalability, recall and accuracy of the approach. We have also addressed the geospatial aspect of queries, notably by using geographical external resources that provide spatial semantic knowledge, in order to deal with heterogeneous geographical locations; existing query propagation approaches focus on the thematic aspects only. In addition, the approach was implemented with Lightweight Coordination Calculus (LCC) a suitable framework for distributed processes such as propagation in networks on the basis of constraints. The experimentation of the approach shows that the three strategies are complementary to reach relevant databases while reducing the number of databases being accessed. Furthermore, it shows that along with a semantic mapping approach, it contributes to a real time semantic interoperability framework for ad hoc network of geospatial databases. In future work, we aim at developing quality measures for strategies, and therefore extending the previous work we have initiated by developing quality measures for semantic mappings in Bakillah et al. 2009. The role of quality measures for strategies will contribute to the development of a global framework for assessing, in real time, the quality of semantic interoperability in geospatial databases of ad hoc networks. The development of quality measures is linked to further investigation of the parameters that influence the performance of strategies, taking into account the space, time and theme aspects of the concepts involved in queries.

CHAPTER 7

SIM-NET: A View-Based Semantic Similarity Model for Ad Hoc Networks of Geospatial Databases

M. BAKILLAH, M.A. MOSTAFAVI, J. BRODEUR, Y. BÉDARD

Published in *Transactions in GIS*, 13(5), 417-447.

7.1 Présentation de l’article

Dans le cadre conceptuel proposé au Chapitre 3, nous avons présenté le modèle de mapping sémantique G-MAP afin de résoudre les hétérogénéités entre les concepts MVAC. Cependant, bien que le G-MAP produise des relations sémantiques qualitatives entre les concepts, une similarité sémantique quantitative est également requise afin de pouvoir distinguer entre les paires de concepts qui sont liées par la même relation sémantique, mais dont le lien sémantique n’a pas nécessairement la même force. Dans ce chapitre, nous présentons une nouvelle mesure de similarité sémantique qui remplit ce rôle. Ce chapitre a été publié en 2009 dans le journal *Transactions in GIS*.

La similarité sémantique est une notion fondamentale dans le domaine des sciences de l’information géographique pour réaliser l’interopérabilité sémantique de données géospatiales. Jusqu’à maintenant, plusieurs modèles de similarité sémantique ont été proposés. Cependant, peu d’entre eux ont été conçus pour s’adapter aux particularités de l’évaluation de la similarité sémantique dans un réseau ad hoc. De plus, plusieurs modèles utilisent une représentation des concepts où les caractéristiques de ces concepts sont considérées comme étant indépendantes. Cette représentation simplifiée réduit la richesse de la représentation des concepts géospatiaux. Dans cet article, nous présentons Sim-Net, un nouveau modèle de similarité sémantique pour les réseaux ad hoc basé sur le langage de la logique descriptive. Sim-Net intègre le paradigme multi-vues. Il permet de représenter les connaissances inférées, c'est-à-dire, les dépendances implicites entre les caractéristiques des concepts, et de les inclure dans le calcul de la similarité. Sim-Net s'appuie sur les concepts de système de référence sémantique et d'analyse formelle des concepts (Formal Concept Analysis – FCA), lesquels sont combinés pour établir un cadre de référence sémantique commun pour les ontologies du réseau ad hoc, appelé View Lattice. Le modèle Sim-Net fait la distinction entre les concepts qui appartiennent au même domaine ou à différents domaines. De plus, il prend en compte le voisinage d'un concept dans le réseau ad hoc. Un exemple d'application est présenté afin de démontrer l'impact positif de Sim-Net.

7.2 Presentation of the Article

In the proposed real time semantic interoperability framework of Chapter 3, we have presented the G-MAP semantic mapping model to resolve heterogeneity among MVAC concept. However, while G-MAP produces qualitative semantic relations between concepts, a quantitative similarity is also required to distinguish between the pairs of concepts that are related by the same type of qualitative relation but that may not be semantically related with the same strength. In this chapter, we present a new semantic similarity measure that fulfills this role. This chapter was published in 2009 as an article in *Transaction in GIS*.

Semantic similarity is a fundamental notion in GIScience for achieving semantic interoperability among geospatial data. Until now, several semantic similarity models have been proposed; however, few of these models address the issues related to the assessment of semantic similarity in ad hoc networks. Also, several models are based on a definition of concepts where features are independent, an assumption that reduces the richness of the geospatial concept representation. This paper presents the conceptual basis for Sim-Net, a novel semantic similarity model for *ad hoc* networks based on Description Logics (DL). Sim-Net is based on the multi-view paradigm. This paradigm is used to include *inferential knowledge* in semantic similarity measurement, where inferential knowledge refers to the knowledge about implicit dependencies between features of concepts. In Sim-Net, assessing semantic similarity relies on the notions of Semantic Reference Systems and

Formal Concept Analysis (FCA), which are combined to establish a common semantic reference frame for ontologies of the *ad hoc* network called the view lattice. The Sim-Net semantic similarity measure distinguishes concepts that belongs to different or similar domains and takes into account the neighbours of a concept in the network. An application example is used to show the positive impact of Sim-Net.

7.3 Introduction

Technological advances have allowed a paradigm shift from isolated information systems to *ad hoc* networks. The GIS community has also taken advantage of these developments, resulting in increasing availability of geospatial data and services. This was meant to fulfill the need of numerous applications to use data from several independent geographical information systems (Lutz et al. 2003; Lemmens 2006), for example in disaster management (Bakillah et al. 2007). However technical developments are not sufficient as we must also resolve semantic discrepancies, i.e. achieve semantic interoperability (Goodchild et al. 1998; Harvey, Kuhn et al. 1999; Kavouras et al. 2005; Bian and Hu 2007). Ontologies, which are explicit specifications of a conceptualization (Gruber 1993), are key components to support semantic interoperability (Brodeur et al. 2003; Fonseca et al. 2005, Arpinar et al. 2006; Fallahi et al. 2008; Kavouras and Kokla 2008), since their role is to make explicit the semantics of data (Kuhn 2003; Agarwal 2005). Ontologies are often used to describe data resources, such as database schemas and contents (Brodaric et al. 2009). Nevertheless, the problem of achieving semantic interoperability is still not resolved since ontologies are semantically heterogeneous. From this point of view, semantic similarity plays a major role for achieving semantic interoperability. It is used to determine if geospatial concepts are close in meaning, so users of different geospatial data sets can exchange data in a meaningful way. However, while ad hoc networks become widespread, the concern of assessing semantic similarity between concepts of ontologies in an ad hoc network has rarely been addressed, except in the non-geospatial domain (Castano et al. 2006). Semantic similarity in ad hoc networks is not the same as semantic similarity between two concepts, because we have to consider that ontologies may describe different domains, and the neighbourhood of a concept may influence similarity. Also, several existing models for comparing concepts consider that features of concepts are independent, an assumption that reduce the expressivity of geospatial concepts (e.g. temperature depends on altitude, geometrical representation depends on descriptive properties, etc.). As stated in Janowicz et al. (2008), semantic similarity depends on the concept representation: a poor description leads to inaccurate results because some factors are not taken into account. This is why we argue that to obtain more accurate semantic similarity results (in terms of what is being compared) we must integrate dependent properties into the definition of concepts.

This paper describes Sim-Net, a novel semantic similarity model for ad hoc networks of geospatial databases that addresses these issues. Sim-Net addresses the requirements posed by an ad hoc network on semantic similarity: first by using a standard knowledge representation language, i.e., Description Logics (DL). The assessment of semantic similarity is supported by a common semantic reference frame which is established using the notion of Semantic Reference Systems proposed by Kuhn (2003) and Formal Concept Analysis (FCA). A contribution of Sim-Net is to explore the logic view paradigm as a mean to include *inferential knowledge* in the semantic similarity measurement. Inferential knowledge allows to discover the implicit relationships among properties of concepts, rather than considering concepts as unstructured sets of independent properties.

This paper is organized as follows: Section 7.4 is a state-of-art on semantic similarity, where we review the different types of semantic similarity models and compare some of their features with Sim-Net. Section 7.5 discusses the requirements for semantic similarity in ad hoc networks of geospatial databases. Section 7.6 presents the conceptual basis of Sim-Net: the view paradigm for representing dependencies between properties and inferential knowledge; the common semantic reference frame on which the assessment of semantic similarity is based; the reasoning method for discovering semantic relations between concepts of different ontologies and the Sim-Net semantic similarity measure. Section 7.7 presents an application example of Sim-Net that illustrates its main features, and Section 7.8 concludes this paper.

7.4 State of the Art on Semantic Similarity

Numerous semantic similarity models have been proposed in the literature; models for the geospatial domain have been described in a recent review by Schwering (2008). The models for comparing concepts include geometric, feature and network models.

Geometric models are based on the concept of multidimensional vector spaces. Each dimension represents a property of concepts (for ex., size); the values of a property (for ex., thin, large) are shown as values on the corresponding dimension. Concepts are represented by multidimensional regions in this vector space. Semantic similarity between concepts can be computed as a function of spatial distance (e.g. Minkowski distance) between vectors forming the boundaries of the regions representing concepts (Schwering and Raubal 2005). This model assumes that the compared concepts are defined with the same dimensions. This is not the case for concepts of different ontologies where a similar real world phenomenon can be represented with different properties that are relevant to the application domain. In Schwering and Kuhn (2009), this model was extended to take into account relations between concepts. In this extended model, concepts are defined with different dimensions; however, dimensions either match or mismatch, but there is no partial match. In Schwering and Raubal (2005) and Schwering and Kuhn's models, properties are independent of each other; however, Raubal (2004) proposed that dependent properties may be modelled via non-orthogonal dimensions, but this idea was not further formalized.

In network models, concepts are nodes in a graph, and their semantics are given by their relative position in this graph (Raftopoulou and Petrakis 2005). Semantic similarity is a decreasing function of the distance between two concepts. Several network models have been proposed, which assign weights to the different types of relationships (Maguitman et al. 2005), combine the shortest path length with the depth first common ancestor concept (Li et al. 2003), compare neighbouring nodes of concepts (Do and Rahm 2002), or incorporate the notion of information content (Resnick 1999). The drawbacks of network models is that they often assume a representation of concepts with labels only, while geospatial concepts are more complex, having spatial, temporal and thematic properties.

While Sim-Net uses notions of network models, it incorporate a complex representation of concepts adapted to the geospatial domain.

Feature models represent concepts as unstructured sets of features; they are based on set theory. The ratio model of Tversky (1977) evaluates the semantic similarity according to the ratio of common and exclusive features. Rodriguez and Egenhofer's Matching Distance model (2003) combines both the ratio model of Tversky with network distance. An example of feature model that computes qualitative relationships among concepts is the geosemantic proximity model (Brodeur and Bédard 2001), which provides geosemantic proximity predicates based on Egenhofer’s topological predicates (Egenhofer 1993). Those feature models cannot provide partial matches between features since features either match or mismatch. However, the Matching Distance model has been extended to allow measuring such partial matches (Bakillah et al. 2006). Nevertheless, these features models remain problematic since features are assumed to be independent.

The geometric, feature and network models are all inspired from the human perception of similarity (Schwering and Kuhn 2009). At another level, we must also mention the logic-based semantic similarity models, which define concepts with a logical language such as Description Logics (DL). For example, d'Amato et al. (2005) proposed a semantic similarity measure for ALC Description Logics. This measure uses instances of concepts. It should be noted that geometric, feature and network models can be represented with Description Logics (Borgida et al. 2005). Another example of logic-based model is Sim-DL by Janowicz (2006). Sim-DL, as the Sim-Net model, is based on Description Logics (DL). Sim-DL compares concepts described with ALCNR DL (a subset of DL). It compares primitive concepts, roles, and cardinality restrictions on roles, and provides a weighted sum of similarities with respect to these features. In comparison, Sim-Net also considers datatype properties, which are required for expressing spatial and temporal properties. A difference between Sim-DL and Sim-Net is that we consider that properties are not independent from each other. In Sim-DL it is proposed that the weights for the different similarity terms can be computed based on probabilistic methods, while we compute weights using domain similarity. As Sim-Net is specifically targeted at ad hoc networks, one difference between Sim-Net and other DL-based models is that it includes the notion of inter-ontology neighbourhood. While the notion of neighbourhood was exploited in the Matching Distance model of Rodriguez and Egenhofer (2003), the neighbourhood of a concept can only include concepts that are close in the same ontology, but cannot include other concepts of the network.

7.5 Semantic Similarity in Ad Hoc Networks

7.5.1 Ad Hoc Networks of Geospatial Databases

An ad hoc network is a network where some of the databases (or "nodes") are made available to a community of users for the duration of a specific need (Fernandez 2007). We assume that each database of the ad hoc network commits to a single ontology. Ontologies play a key role by capturing the shared conceptualization of a community of users. Consequently, they support interoperability between different databases (Smith and Mark 1998; Fonseca *et al*. 2005). Figure 7.1 shows how we represent the ad hoc network.

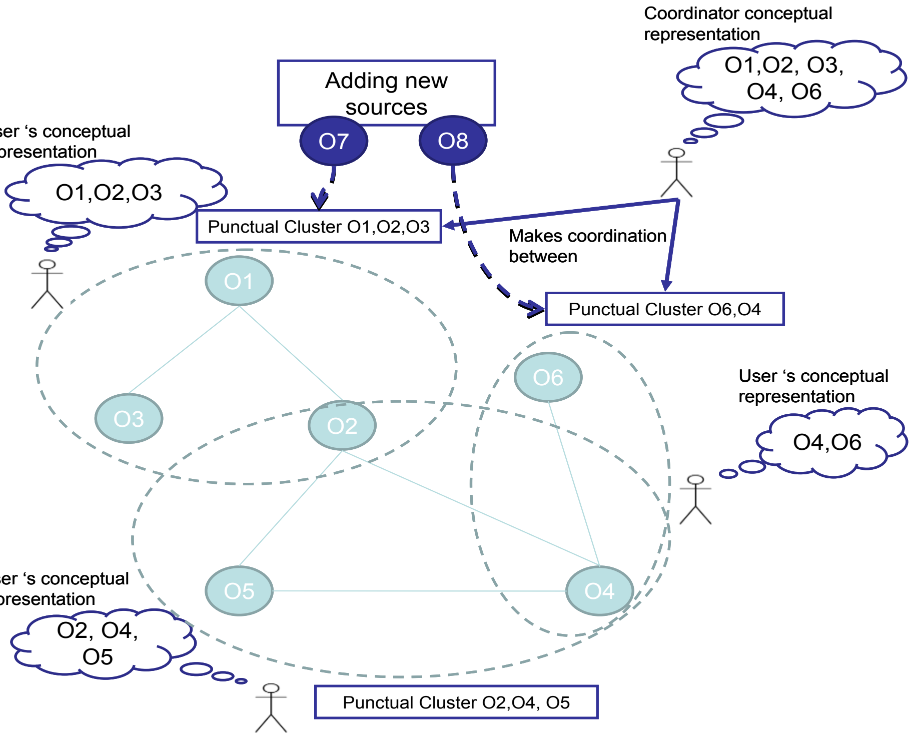

Coordinator conceptual representation
O1,O2, O3, O4, O6
Adding new sources
O7
O8
User 's conceptual representation
O1,O2,O3
Punctual Cluster O1,O2,O3
Makes coordination between
Punctual Cluster O6,O4
O1
O3
O2
O6
User 's conceptual representation
O4,O6
O5
O4
User 's conceptual representation
O2, O4, O5
Punctual Cluster O2,O4, O5

Figure 7.1 Ad hoc network of geospatial databases

Each node of the network (labelled O_i) represents an ontology. The ontologies are gathered in subsets (represented with dotted circles) called *punctual clusters*. Ontologies contained in a punctual cluster are available to a community of users having common interests. They stand for the users' conceptual representation. The ad hoc network is dynamically modified when the users' interest changes or when nodes are added or dropped from the network. When such changes occur, punctual clusters have to be re-organized by a coordinator user. Semantic similarity among concepts of different ontologies can be affected by such changes, since we assume that the neighbourhood of a concept (the other concepts to which it is semantically linked) contribute to define its semantics.

7.5.2 What do we need for Semantic Similarity in Ad Hoc Networks of Geospatial Databases?

- *Anchoring semantic similarity in standard knowledge representation language.* According to Janowicz (2006), the Web Ontology Language (OWL), which is based on DL, is the most widely adopted ontology language for geo-ontologies as it is recommended by W3C (Baader et al. 2003). A semantic similarity model suitable for networks should be based on such commonly accepted language in order to avoid the problem of incompatibility between knowledge representation and comparison criteria (Janowicz 2006). The advantages of DL are its sound semantics and its reasoning capabilities. It also supports complex concept description (Borgida et al. 2005). However, while very expressive families of DL exist, they are not appropriate because of their complexity, which makes them undecidable.

- *Defining a common semantic reference frame for the ad hoc network.* Assessing semantic similarity among concepts from different ontologies in the ad hoc network requires that the concepts can be referenced in a common semantic reference system (Kuhn 2003), which is analogous to spatial reference systems. Semantic reference systems are more than ontologies, but ontologies are a core component of them (Kuhn and Raubal 2003). The semantic datum's role is to ground the meaning of basic terms. The semantic reference frame is the formally-defined framework to which terms can be related to obtain meaning (it could be a top-level ontology). Semantic referencing is the process of linking the terms of a local model to an element of the semantic reference frame.

- *Determining semantic similarity among concepts of different domains*. An ad hoc network is populated with ontologies that describe different domains. Several semantic similarity models are based on the assumption that similar concepts have similar properties. In an ad hoc network, we cannot always make this assumption, since ontologies of different domains describe the same concept with different properties.

- *Representation of concepts suitable for geospatial concepts*. Geospatial concepts are often related through logic rules that express dependencies between their properties. There are several reasons which demonstrate why logic rules are fundamental elements. First, geospatial concepts are often described with physical properties (e.g. temperature, altitude, size, density, etc.); by nature, physical properties tend to depend on each other (e.g. temperature depends on altitude). Also, the geometry and temporality used to represent geospatial objects depends on other spatial, temporal or thematic properties, for example, watercourses larger than 7,5 are represented by surfaces while those that are thinner than 7,5 m are represented by lines.

- *Propagation of semantic similarity in ad hoc networks.* Semantic similarity can be deduced from existing relationships between concepts, through inference mechanisms based, for example, on the transitivity property of semantic relationships. Transitivity means that if a first concept C1 is similar to C2, and C2 is similar to C3, than C1 is similar to C3. However, this property has been criticized on the basis of the famous example of James: a lamp is similar to the moon and the moon is similar to a ball; but a lamp is not similar to a ball. This is because similarity depends on what is being compared. Lamp and moon are similar with respect to their function (to provide light), while moon and ball are similar with respect to their shape. Therefore, we should be careful when using transitivity and ensure that we compare concepts with respect to the same aspect. In this paper, however, we do not address this requirement and leave it for future work.

7.6 Sim-Net Semantic Similarity Model

The conceptual basis for Sim-Net consists of the following elements: first we present the logic view paradigm (section 7.4.1). We use the Formal Concept Analysis (FCA) theory to build a common semantic reference frame to which concepts and views of local ontologies can be referenced (section 7.4.2). In section 7.4.3, we give the DL-based reasoning rules used by Sim-Net to determine the semantic relationships among views and among concepts. Finally, we give the Sim-Net semantic similarity measure.

7.6.1 The Logic View Paradigm and View Extraction Method

Several researchers have shown interest in the view paradigm, both in the database field (Debrauwer 1998; Bédard and Bernier 2002; Benchikha et al. 2005; Parent et al. 2006) and in the domain of ontologies (Noy and Musen 2004; Bhatt et al. 2006; Stuckenschmidt 2006; Wouters et al. 2008). In the database field, views are used to handle multi-representation (Bédard and Bernier 2002; Parent et al. 2006). Views can also represent the different states of an evolving object (Debrauwer 1998). In the ontology domain, the view paradigm supports ontology reuse by selecting only parts of an ontology that are relevant in a given context. In our approach, views are also used to handle inferential knowledge obtained from logic rules.

Geospatial concepts are complex since they are described by spatial properties such as shape and position, spatial relations (Schwering 2008) and temporal relations. Furthermore, they are often described by logic rules that constraint their property values, for example when an *industry* has for property value *type of product = toxic substance*, it must be situated at more than 3km from *residential areas*. The knowledge extracted from the conjunction of these logic rules is called *inferential knowledge* (Steffens 2005). For example, consider a concept "road" with properties road type ={street, boulevard} and "number of lanes". The following logic rule expresses a relation between "road type" and "number of lane": (road type(X) = boulevard)→(number of lanes(X) ≥2). If we have another rule (number of lanes(X)≥2)→(road geometry = multi-lines), we can infer a new relationship between "road type" and "road geometry", in the form of a new logic rule: (road type(X) = boulevard)→(road geometry = multi-lines). The general form of rules is:

$$r:[p_i(X)=v_i] \rightarrow [p_j(Y)=v_j] \tag{1}$$

where v_i is an element of the range of property p_i, v_j is an element of the range of property p_j, and X and Y are variables of instances of concepts. The first member of the logic rule is called the antecedent and the second member is the consequent.

The view paradigm consists in expressing inferential knowledge obtained from logic rules with logic views of a concept. From the above example, we see that a logic view of the concept road can be "boulevard" with properties road geometry = multi-lines and number of lanes(X) ≥2. Because of the lack of space, we are only giving an overview of the view extraction method. We assume that a concept is defined by its name, a set of properties from the categories shown in Figure 7.2, a set of relationships and a set of logic rules. Figure 7.2 shows a classification of seven sub-types of properties related with "is-a-kind-of" relations.

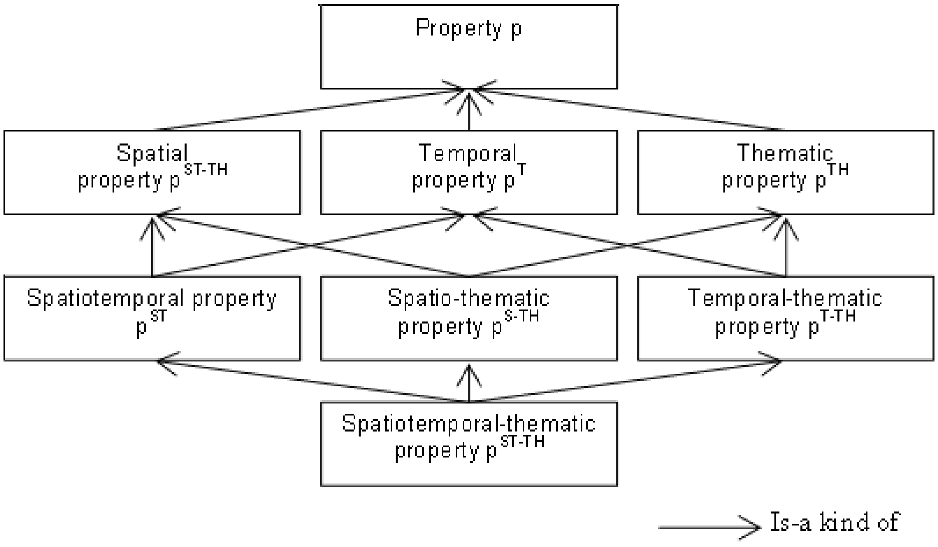

Property p
Spatial property p^{ST-TH}
Temporal property p^{T}
Thematic property p^{TH}
Spatiotemporal property p^{ST}
Spatio-thematic property p^{S-TH}
Temporal-thematic property p^{T-TH}
Spatiotemporal-thematic property p^{ST-TH}
Is-a kind of

Figure 7.2 The different types of properties of concepts

Spatial properties are properties whose range is a spatial datatype (point, line, polygon, etc.). The range of temporal properties is a temporal datatype (instant, period), while thematic properties have string or numerical values as a range. The is-a-kind-of relation indicates that sub-properties inherit the characteristics of the super-properties. For example, spatiotemporal properties inherit the characteristics of spatial and temporal properties, so their range is a spatial datatype associated to a temporal datatype.

The idea behind the view extraction method is to use each logic rule to extract a partial view of the concept, and then combine the partial views with compatible values of properties. The view extraction method is summarized as follows:

1. Extracting existing logic rules from the definition of concepts. Two cases can be considered: either rules are explicit ontology elements that were defined at design time: in this case, we can obtain them by accessing the ontology. Otherwise, it may be that such rules are not directly available in the ontology; in this case, they could be discovered with association rule mining techniques.

2. Applying the inference mechanism between existing rules to discover new rules. This mechanism states that if the consequent of a first logic rule *r1* implies the antecedent of a second logic rule *r2,* then the antecedent of *r1* also implies the consequent of *r2*:

$$\begin{array}{l}(r1:[p_i(X)=v_i]\rightarrow[p_j(Y)=v_j])\wedge\\(r2:[p_j(Y)=v_l]\rightarrow[p_k(Z)=v_k])\wedge\\([p_j(Y)=v_j]\rightarrow[p_j(Y)=v_l])\rightarrow\\(r3:[p_i(X)=v_i]\rightarrow[p_k(Z)=v_k])\end{array}\qquad(2)$$

3. Create a partial view from each logic rule, following the association of properties and property values stated by the rule.

4. Merge partial views to obtain the views of the concept: we merge all partial views that contain compatible values of properties, until the view specifies the range of each property of the concept. The validity of extracted views can be verified through consistency checking (verifying that views are instanciable).

Once views are extracted from concepts, we obtain a set of ontologies extended with views. The next step is to reference the concepts and these views into a common semantic reference frame.

7.6.2 Building the Common Semantic Reference Frame: The View Lattice

We use the notions of reference frame and referencing described in the theory of Semantic Reference Systems (Kuhn 2003). The view lattice plays the role of the semantic reference frame, and it is built using the Formal Concept Analysis (FCA) method. FCA has been used in previous semantic interoperability approaches (Bian and Hu 2007; Curé and Jeansoulin 2009). It provides a framework for placing concepts of different ontologies in a single hierarchy called the Concept Lattice (Ganter and Wille 1999). A view lattice is a set of formal concepts and formal views that are linked by inheritance relationships. We summarize the method for building the view lattice as follows:

Step 1) **Generation of the set of reference concepts** – reference concepts define the common vocabulary for a set of local ontologies. Our approach to generate the reference concepts is as follows: first, we gather all concepts of local ontologies. We group the concepts that are synonyms into subsets. These subsets are the reference concepts. Then, we find the subsets that contain concepts related with is-a relations and add this knowledge to the definition of reference concepts. Synonyms and is-a relations can be identified with WordNet (Miller 1995), a domain-independent thesaurus for the English language. Building this common vocabulary solves lexical heterogeneity across ontologies.

Step 2) **Projection of local concepts to reference concepts** – The next step is to project the concepts from different ontologies (local concepts) to the reference concepts in a projection matrix. Table 7.1 shows an example.

Table 7.1 Example of projection of local concept to reference concepts

Local Concepts	Reference Concepts				
	(Stream/	(City/	Land	(Lowland)	(Upland)

	Watercourse)	Urban Area)		is-a Land	is-a Land
O1: Stream	x				
O2: Watercourse	x				
O1: City		x			
O2: Urban Area		x			
O1: Land			x	x	x
O2: Lowland				x	
O2: Upland					x

Each local concept is identified by a prefix (in this example, O1 or O2) that identifies the local ontology it belongs to. An *x* sign indicates when a local concept contains a reference concept.

Step 3) **Identification of formal concepts from the projection matrix** – In the original FCA theory, a formal concept is a pair (A = set of objects, B = set of attributes), where objects in A are described in terms of attributes of B. In our context, it is local concepts that are described in terms of reference concepts. Therefore, we apply the FCA theory by establishing the correspondences object-local concept and attribute-reference concepts. A formal concept is a pair *FC* = <(local concepts), (reference concepts)> that co-occur in the matrix, for example: <(Stream, Watercourse),(Stream/Watercourse)> and <(Land, Lowland), (Lowland)>.

Step 4) **Identification of inheritance relationships and generation of the lattice** – We verify whether a formal concept includes another formal concept in terms of reference concept, and build the upper part of the view lattice, which contains all formal concepts (Figure 7.3):

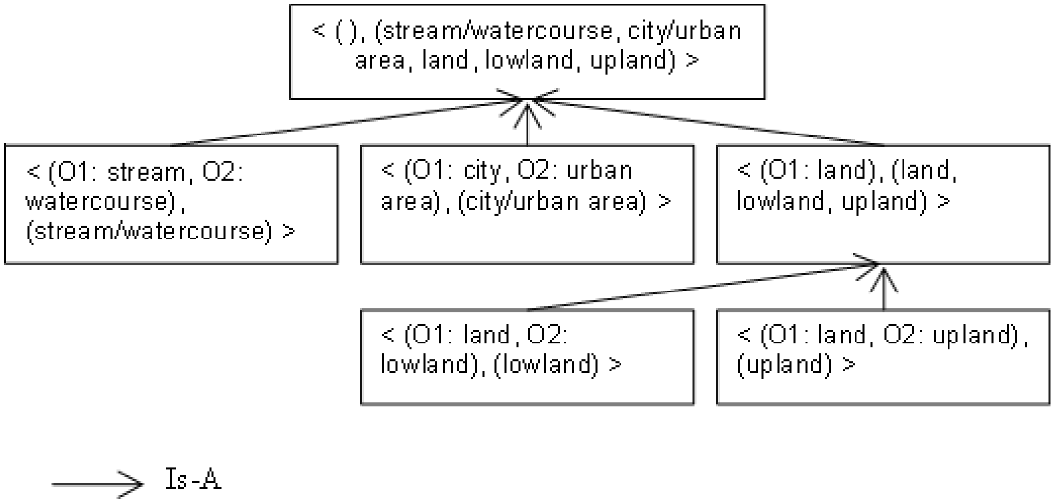

< ( ), (stream/watercourse, city/urban area, land, lowland, upland) >
< (O1: stream, O2: watercourse), (stream/watercourse) >
< (O1: city, O2: urban area), (city/urban area) >
< (O1: land), (land, lowland, upland) >
< (O1: land, O2: lowland), (lowland) >
< (O1: land, O2: upland), (upland) >
Is-A

Figure 7.3 Upper part of the view lattice containing only formal concepts

Step 5) **Expansion of the upper part of the view lattice with formal views** – This is similar to steps 1 to 4, however the local views are projected to reference properties and values of properties. Table 7.2 illustrates an example of projection of the views of concepts Stream and Watercourse, where an x sign indicates that a view has a reference property and value.

We identify the formal views, which are pairs *FC* = <(local views), (reference property: reference value)> that co-occur in the matrix. The formal views for Table 7.2 are the following, and the complete view lattice that expands the upper part of Figure 7.3 is shown in Figure 7.4.

FV_1 = <(V_1(O1:Stream), V_1(O2:Watercourse)), (Stream Geometry/Watercourse Geometry: Surface/Region)>

FV_2 = <(V_2(O1:Stream), V_1(O2: Watercourse)), (Stream Class/Watercourse Category: River)>

FV_3 = <(V_1(O1:Stream), (Stream Class/Watercourse Category: Canal; Stream Geometry/Watercourse Geometry: Surface/Region)>

FV_4 = <(V_2(O1:Stream), (Stream Class/Watercourse Category: River; Stream Geometry/Watercourse Geometry: Surface/Region →Time interval)>

FV_5 = <(V_1(O2:Watercourse), (Stream Class/Watercourse Category: River; Stream Geometry/Watercourse Geometry: Surface/Region)>

FV_6 = <(V_2(O2:Watercourse), (Stream Class/Watercourse Category: Ditch; Stream Geometry/Watercourse Geometry: Line)>

Table 7.2 Example of projection of local views to reference properties and values

Local views	Reference properties and values					
	Stream Class/Watercourse Category			Stream Geometry/Watercourse Geometry		
	Canal	River	Ditch	Surface/Region	Surface/Region →Time interval	Line
V_1 (O1:Stream)	x			x		
V_2 (O1:Stream)		x			x	
V_1 (O2: Watercourse)		x		x		
V_2 (O2: Watercourse)			x			x

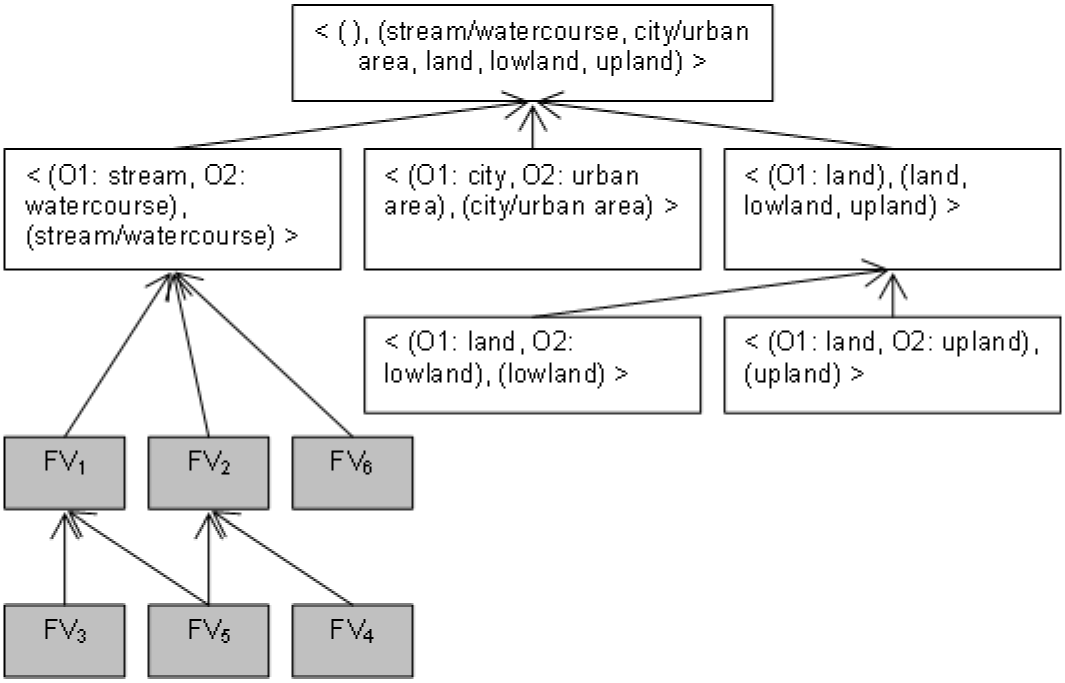

< ( ), (stream/watercourse, city/urban area, land, lowland, upland) >
< (O1: stream, O2: watercourse), (stream/watercourse) >
< (O1: city, O2: urban area), (city/urban area) >
< (O1: land), (land, lowland, upland) >
< (O1: land, O2: lowland), (lowland) >
< (O1: land, O2: upland), (upland) >
FV1
FV2
FV6
FV3
FV5
FV4

Figure 7.4 Complete view lattice with formal concepts (white boxes) and formal views (grey boxes)

Step 6) **Referencing Views and Concepts to the View Lattice** – A concept C is referenced to a formal concept FC (denoted $C \xrightarrow{\text{RefTo}} FC$) if it verifies the following condition: the set of local concepts of FC is the smallest set that contains C in the view lattice.

A view V is referenced to a formal view FV (denoted $V \xrightarrow{\text{RefTo}} FV$) if it verifies (V1) to (V3):

(V1) Properties of FV are all included in properties of V;

(V2) For all properties owned both by FV and V, the associated property values in V are all included in the set of property values of FV;

(V3) FV is the lowest formal view of the view lattice that satisfies conditions (V1) and (V2).

Once views and concepts of different ontologies are referenced to the common view lattice, the setting is ready for assessing semantic similarity with Sim-Net.

7.6.3 The Sim-Net Semantic Similarity Model

Sim-Net is a DL-based semantic similarity model for *ad hoc* networks that determines semantic relationships among concepts and their semantic similarity value. Consider a set of N ontologies {O_1, O_2 ... O_N} from the *ad hoc* network. Sim-Net takes this set of ontologies and returns a set of pairs of concepts, O_i: C_j and O_k: C_l, their associated semantic relationship R and semantic similarity value SN:

$$Sim-Net : O_1 \times O_2 ... O_N \rightarrow \left\langle O_i : C_j ; O_k : C_l ; R(C_j, C_l) ; SN(C_j, C_l) \right\rangle \qquad (3)$$

Sim-Net is different from other semantic similarity model because it is based on the following assumptions:

- Ontologies in ad hoc networks describe different domains; therefore, we cannot assume that concepts describing the same reality necessarily share common properties;

- When a concept from a first ontology is linked to concepts of other ontologies, these concepts also contribute in defining its semantics. Therefore assessing semantic similarities between concepts of two ontologies is different than between ontologies of a network.

In the following, we provide background information on DL and we explain how semantic relationships and semantic similarity values are determined.

7.6.3.1 Description Logics

Description Logics are a family of representation languages widely adopted for knowledge representation and reasoning (Baader *et al.* 2003, Lemmens 2006, Fallahi *et al.* 2008). They are based on the notion of concepts and roles. Constructors (universal quantification, existential restriction, conjunction, etc.) allow defining complex concepts and complex roles from primitive ones. Common constructors are listed in Table 7.3.

Table 7.3 Syntax and semantics of Common Description Logic Constructors

Name	**Syntax**	**Semantic**
Top concept		Δ^I
Bottom concept		$\emptyset$
Atomic concept	C	$C^I \subseteq \Delta^I$
Atomic role	R	$\mathrm{R}^\mathrm{I} \subseteq \Delta^I \times \Delta^I$
Full negation	$\neg C$	Δ^I / C^I
Concept equality	$C \equiv D$	$C^I = D^I$
Concept inclusion	$C \subseteq D$	$C^I \subseteq D^I$
Concept union	$C \cup D$	$C^I \cup D^I$
Concept intersection	$C \cap D$	$C^I \cap D^I$
Role equality	$R \equiv S$	$R^I = S^I$
Role inclusion	$R \subseteq S$	$R^I \subseteq S^I$
Existential quantification	$\exists R.C$	$\{a \in \Delta^I \mid \exists b.(a,b) \in R^I \wedge y \in C^I\}$
Value restriction	$\forall R.C$	$\{a \in \Delta^I \forall b.(a,b) \in R^I \rightarrow y \in C^I\}$
Maximum number restriction	$\leq NR.C$	$\{a \in \Delta^I \mid \{b \in \Delta^I \mid (a,b) \in R^I \wedge b \in C^I\} \mid \leq n\}$
Minimum number restriction	$\geq NR.C$	$\{a \in \Delta^I \mid\mid \{b \in \Delta^I \mid (a,b) \in R^I \wedge b \in C^I\} \mid \geq n\}$

The semantics of those constructors are given by an interpretation $\mathrm{I}=(\Delta^\mathrm{I}, {}^\mathrm{I})$, where Δ^I is the set of instances and ${}^\mathrm{I}$ is the function that associate instances to their concepts. The different forms of description logics are determined by the constructors that are used, and give the expressive power of DL. Sim-Net is based on the *SHIQ(D)* DL form (where *SHIQ(D)* is a subset of DL that allows for the expression of inverse roles, qualified number restriction, and datatypes properties); this choice is motivated by the fact that we need to be able to express inverse roles (*I*) which are necessary to describe spatial relationships (e.g. IsIncludedIn is the inverse of Includes), qualified number restriction (*Q*) for expressing cardinality restrictions on relationships, and datatype properties (*D*), for example HasArea, that link a concept with a spatial datatype such as polygon:

$$Lake \equiv Waterbody \cap \exists HasArea.Polygon$$

7.6.3.2 Reasoning with DL for Determining Semantic Relationships in Ad Hoc Network

Sim-Net uses reasoning rules for determining semantic relationships at two levels: between views of different concepts, and between concepts. Consider the DL-expression of views O_1: V_1 and O_2: V_2 which are given by:

$$O_1 : V_1 \equiv \bigcap_{A_i^V \in \{\text{Primitive views}\}} \quad \bigcap_{R_j \in \{\text{Primitive Role}\}} \quad \bigcap_{T_k \in \{\text{Datatype Property}\}} \qquad (4)$$

with *i* element of [1, 2, *i*’], *j* element of [1, 2, *j*’] and k element of [1, 2, *k*’].

$$O_2 : V_2 \equiv \bigcap_{B_m^V \in \{\text{Primitive views}\}} \quad \bigcap_{S_n \in \{\text{PrimitiveRole}\}} \quad \bigcap_{U_p \in \{\text{Datatype Property}\}} \qquad (5)$$

with *m* element of [1, 2, *m*’], *n* element of [1, 2, *n*’] and p element of [1, 2, *p*’]. Each view is defined by the conjunction of a set of primitive views (A_i^V, B_m^V), a set of primitive roles (R_j, S_n), and a set of datatype properties (T_k, U_p). Each view is referenced to a formal view of the view lattice *L*:

$$O_1 : V_1 \xrightarrow{\text{RefTo}} L : FV_1 \text{ and } O_2 : V_2 \xrightarrow{\text{RefTo}} L : FV_2 .$$

Concepts are defined as sets of views, and their definition is given by:

$$O_1 : C_1 \equiv V_1^1 \cup V_1^2 \cup .... \cup V_1^i ... \cup V_i^N \tag{6}$$

$$O_2 : C_2 \equiv V_2^1 \cup V_2^2 \cup .... \cup V_2^j ... \cup V_2^M \tag{7}$$

Each concept is also referenced to a formal concept of the common view lattice L:

$$O_1 : C_1 \xrightarrow{\text{RefTo}} L : FC_1 \text{ and } O_2 : C_2 \xrightarrow{\text{RefTo}} L : FC_2 .$$

Sim-Net computes semantic relationships among views of concepts with view-reasoning rules that define the conditions for a semantic relationship to be verified (second row of Table 7.4). Then, Sim-Net deduces the relationship between two concepts by reasoning with the relationships between their views, using reasoning rules defined in the third row of Table 7.4. The principle for deducing relationships between concepts from relationships between views is that each concept is a set of views. The semantic relationships are based on classical set theory: *equivalence*, *generalisation*, *overlap* and *disjointness*. In addition, we consider semantic relationships that can be established between concepts (views) of different domains. Classically, equivalence between two concepts (views) is established only if they have exactly the same set of properties and relationships. In different domains, two concepts representing the same reality may have different properties. In this case, we allow two concepts (views) to be *cross-domain equivalent* if they have different properties but are related to the same formal concept (formal view) in the view lattice. We apply the same reasoning for defining the other cross-domain relationships.

Table 7.4 Semantic relationships and reasoning rules

SEMANTIC RELATIONSHIPS	VIEWS REASONING RULES	CONCEPT REASONING RULES
STRONG EQUIVALENCE	**EXPRESSION:** $O_1 : V_1 \xrightarrow{\equiv} O_2 : V_2$ **RULES:** 1) $\forall A_i^V , i = 1...i', \exists B_l^V$ where $A_i^V \equiv B_l^V$ and	**EXPRESSION:** $O_1 : C_1 \xrightarrow{\equiv} O_2 : C_2$ **RULES:** 1) $\forall V_1^i , 1 \le i \le N, \exists V_2^j , 1 \le j \le M$ where $V_1^i \xrightarrow{\equiv} V_2^j$ 2)

	$\forall B_l^V, l=1...l', \exists A_i^V$ where $A_i \equiv B_l$ 2) $\forall R_j, j=1...j', \exists S_m$ where $R_j \equiv S_m$ and $C_j^V \equiv C_m^V$ and $\forall S_m, m=1...m', \exists R_j$ where $S_m \equiv R_j$ and $C_m^V \equiv C_j^V$ 3) $\forall T_k, k=1...k', \exists U_p$ where $T_k \equiv U_p$ and $F_k \equiv G_p$ and $\forall U_p, p=1...p', \exists T_k$ where $U_p \equiv T_k$ and $G_p \equiv F_k$	$\forall V_2^j, 1 \le j \le M, \exists V_1^i, 1 \le i \le N$ where $V_2^j \xrightarrow{\equiv} V_1^i$
CROSS-DOMAIN EQUIVALENCE	**EXPRESSION:** $O_1 : V_1 \xrightarrow{\equiv^*} O_2 : V_2$ **RULES:** $O_1 : V_1 \xrightarrow{\text{RefTo}} L : FV_1$ and $O_2 : V_2 \xrightarrow{\text{RefTo}} L : FV_2$ with $FV_1 \equiv FV_2$	**EXPRESSION:** $O_1 : C_1 \xrightarrow{\equiv^*} O_2 : C_2$ **RULES:** $O_1 : C_1 \xrightarrow{\text{RefTo}} L : FC_1$ and $O_2 : C_2 \xrightarrow{\text{RefTo}} L : FC_2$ with $FC_1 \equiv FC_2$, or 1) $\forall V_1^i, 1 \le i \le N, \exists V_2^j, 1 \le j \le M$ where $V_1^i \xrightarrow{\equiv^*} V_2^j$ 2) $\forall V_2^j, 1 \le j \le M, \exists V_1^i, 1 \le i \le N$ where $V_2^j \xrightarrow{\equiv^*} V_1^i$
STRONG GENERALIZATION (INVERSE OF STRONG SPECIALIZATION)	**EXPRESSION:** $O_1 : V_1 \xrightarrow{\supset} O_2 : V_2$ **RULES:** 1) $\forall A_i^V, i=1...i', \exists B_l^V$ where $A_i^V \equiv B_l^V$ 2) $\forall R_j, j=1...j', \exists S_m$ where $R_j \equiv S_m$ and $C_j^V \equiv C_m^V$ or $C_j^V \supset C_m^V$	**EXPRESSION:** $O_1 : C_1 \xrightarrow{\supset} O_2 : C_2$ **RULES:** 1) $\forall V_2^j, 1 \le j \le M, \exists V_1^i, 1 \le i \le N$ where $V_1^i \xrightarrow{\equiv} V_2^j$ or $V_1^i \xrightarrow{\supset} V_2^j$ 2) $\exists V_1^i, 1 \le i \le N$ where

	3) $\forall T_k, k = 1...k', \exists U_p$ where $T_k \equiv U_p$ and $F_k \equiv G_p$ or $G_p \subset F_k$	$V_1^i \xrightarrow{\neg\equiv} V_2^j$ and $V_1^i \xrightarrow{\neg\supset} V_2^j$
CROSS-DOMAIN GENERALIZATION (INVERSE OF CROSS-DOMAIN GENERALIZATION)	**EXPRESSION:** $O_1 : V_1 \xrightarrow{\supset^*} O_2 : V_2$ **RULES:** $O_1 : V_1 \xrightarrow{\text{RefTo}} L : FV_1$ and $O_2 : V_2 \xrightarrow{\text{RefTo}} L : FV_2$ with $FV_1 \succ FV_2$	**EXPRESSION:** $O_1 : C_1 \xrightarrow{\supset^*} O_2 : C_2$ **RULES:** $O_1 : C_1 \xrightarrow{\text{RefTo}} L : FC_1$ and $O_2 : C_2 \xrightarrow{\text{RefTo}} L : FC_2$ with $FC_1 \succ$ _ 1) $\forall V_2^j, 1 \le j \le M, \exists V_1^i, 1 \le i \le N$ where $V_1^i \xrightarrow{\equiv^*} V_2^j$, or $V_1^i \xrightarrow{\supset^*} V_2^j$ 2) $\exists V_1^i, 1 \le i \le N$ where $\neg(V_1^i \xrightarrow{\{\equiv, \equiv^*, \supseteq, \supseteq^*\}} V_2^j)$
WEAK OVERLAP	**EXPRESSION:** $O_1 : V_1 \xrightarrow{\cap^w} O_2 : V_2$ **RULES:** 1) $\exists A_i^V, i = 1...i'$ and $\exists B_l^V, l = 1...l'$ where $A_i \equiv B_l$ and/or 2) $\exists S_m, m = 1...m'$ and $\exists R_j, j = 1...j'$ where $R_j \equiv S_m$ and C_j^V and C_m^V are not disjoint and/or 3) $\exists U_p, p = 1...p'$ and $\exists T_k, k = 1...k'$ $T_k \equiv U_p$ and F_k and G_p are not disjoint	**EXPRESSION:** $O_1 : C_1 \xrightarrow{\cap^w} O_2 : C_2$ **RULES:** 1) $\exists V_1^i, 1 \le i \le N, \exists V_2^j, 1 \le j \le M$ where $V_1^i \xrightarrow{\neg\perp} V_2^j$ 2) $\exists V_1^i, 1 \le i \le N$ where $V_1^i \xrightarrow{\perp} V_2^j$ with $1 \le j \le M$ 3) $\exists V_2^j, 1 \le j \le M$ where $V_2^j \xrightarrow{\perp} V_1^i$ with $1 \le i \le N$
STRONG OVERLAP	**EXPRESSION:** $O_1 : V_1 \xrightarrow{\cap} O_2 : V_2$ **RULES:**	**EXPRESSION:** $O_1 : C_1 \xrightarrow{\cap} O_2 : C_2$ **RULES:**

	1) $\exists A_i^V, i = 1...i'$ and $\exists B_l^V, l = 1...l'$ where $A_i \equiv B_l$ and/or $\exists S_m, m = 1...m'$ and $\exists R_j, j = 1...j'$ where $R_j \equiv S_m$ and C_j^V and C_m^V are not disjoint and/or $\exists U_p, p = 1...p'$ and $\exists T_k, k = 1...k'$ $T_k \equiv U_p$ and F_k and G_p are not disjoint 2) $O_1 : V_1 \subset O_a : V$ and $O_2 : V_2 \subset O_a : V$	1) $\exists V_1^i, 1 \le i \le N, \exists V_2^j, 1 \le j \le M$ where $V_1^i \xrightarrow{\neg\perp} V_2^j$ 2) $\exists V_1^i, 1 \le i \le N$ where $V_1^i \xrightarrow{\perp} V_2^j$ with $1 \le j \le M$ 3) $\exists V_2^j, 1 \le j \le M$ where $V_2^j \xrightarrow{\perp} V_1^i$ with $1 \le i \le N$ 4) $\exists O_a : V$ where $O_1 : V_1^i \subset O_a : V$ and $O_2 : V_2^j \subset O_a : V$
CROSS-DOMAIN OVERLAP	**EXPRESSION :** $O_1 : V_1 \xrightarrow{\cap^*} O_2 : V_2$ **RULES:** $O_1 : V_1 \xrightarrow{\text{RefTo}} L : FV_1$ and $O_2 : V_2 \xrightarrow{\text{RefTo}} L : FV_2$ with $FV_1 \prec FV_a, FV_2 \prec FV_a$	**EXPRESSION:** $O_1 : C_1 \xrightarrow{\cap^*} O_2 : C_2$ **RULES:** $O_1 : C_1 \xrightarrow{\text{RefTo}} L : FC_1$ and $O_2 : C_2 \xrightarrow{\text{RefTo}} L : FC_2$ with $FC_1 \prec FC_a, FC_2 \prec$ 1) $\exists V_2^j, 1 \le j \le M, \exists V_1^i, 1 \le i \le N$ where $O_1 : V_1^i \xrightarrow{\text{RefTo}} L : FV_1$ and $O_2 : V_2^j \xrightarrow{\text{RefTo}} L : FV_2$ with $FV_1 \prec FV_a, FV_2 \prec FV_a$
DISJOINTNESS	**EXPRESSION:** $O_1 : V_1 \xrightarrow{\perp} O_2 : V_2$ **RULES:** 1) $\neg\exists A_i, B_l$ where $A_i \equiv B_l$ 2 $\neg\exists R_j, S_m$ where $R_j \equiv S_m$ 3) $\neg\exists T_k, U_p$ where $T_k \equiv U_p$ 4) $\neg\exists O_a : V$ where $O_1 : V_1 \subset O_a : V$ and $O_2 : V_2 \subset O_a : V$	**EXPRESSION:** $O_1 : C_1 \xrightarrow{\perp} O_2 : C_2$ **RULES:** 1) $\neg\exists V_1^i, V_2^j$ where $V_1^i \xrightarrow{\neg\perp} V_2^j$ 2) $\neg\exists FC \neq \mathsf{T}$ where $O_1 : C_1 \xrightarrow{\text{RefTo}} L : FC_1$ and $O_2 : C_2 \xrightarrow{\text{RefTo}} L : FC_2$ with $FC \succ$ 1 $FC \succ FC_2$

The result obtained by applying the above reasoning rules is a single ontology graph which links views and concepts of a set of ontologies (Figure 7.5).

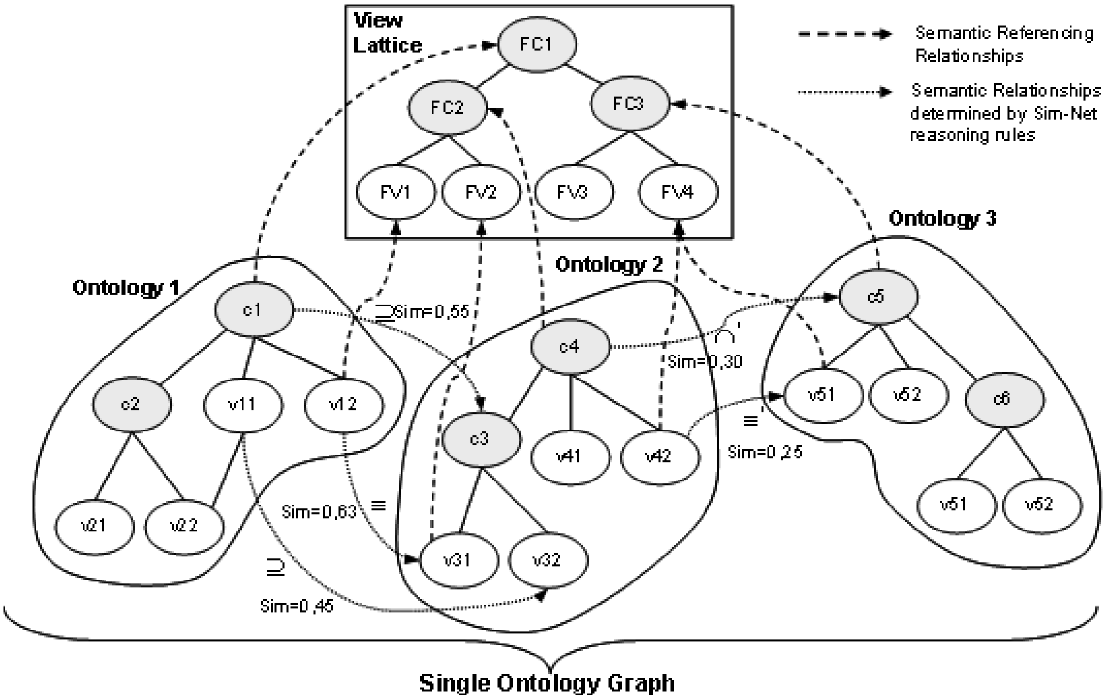


Figure 7.5 The result of finding semantic relationships is a single ontology graph

7.6.3.3 Measuring Semantic Similarity with Sim-Net

The semantic similarity among concepts is defined by a combination function that merges the results of three semantic similarity measures: similarity between views of concepts (Sim_{view}), cross-domain similarity (Sim_{cd}) and inter-ontology neighbourhood similarity (Sim_{ion}):

$$SN_C(C_1,C_2)=\omega_{view}Sim_{view}(C_1,C_2)+\omega_{cd}\left[\frac{Sim_{cd}(C_1,C_2)+Sim_{ion}(C_1,C_2)}{2}\right] \quad (8)$$

The Sim_{view} term accounts for the principle that concepts that have similar views are also similar. The Sim_{cd} term accounts for the principle that concepts are similar if they are referenced to similar formal concepts in the view lattice. This term allows us to find non-zero similarity even if concepts have no common properties, as it can produce when ontologies describe different domains. The Sim_{ion} term accounts for the fact that concepts are similar if their inter-ontology neighbourhoods contain similar concepts. For concepts of similar domains, we expect that Sim_{view} should be more important than Sim_{cd} and Sim_{ion}. Therefore, more weight is given to Sim_{view} if domains of concepts are similar, while more weight is given to Sim_{cd} and Sim_{ion} when domains of concepts are dissimilar. We define latter the computation of weights; however we will first discuss the semantic similarity terms.

7.6.3.3.1 Adapting the Normalized Google Distance to the Ad Hoc Network

The Normalized Google Distance (NGD) was introduced by Cilibrasi and Vitanyi (2007) for assessing semantic distance between concepts in the Web. The semantics of a concept is given by the set of Web pages returned by the Google search engine when this concept is used as the query word. For concepts x and y, the NGD is:

$$\text{N}GD(x,y)=\frac{\max\{\log f(x),\log f(y)\}-\log f(x,y)}{\log M-\min\{\log f(x),\log f(y)\}} \tag{9}$$

f(*x*) is the number of web pages containing x, *f*(x,y) is the number of pages containing x and y and M is the number of pages indexed by Google. NGD is a measure of the probability of co-occurrence of x and y in the Web. The more the concepts will co-occur in the same page, the smaller the semantic distance. The Web is considered as a giant ontology graph. The structure of an *ad ho*c network, where concepts of ontologies are related with semantic relationships discovered by Sim-Net, is similar to the ontology defined by the Web. For each similarity term Sim_{view}, Sim_{cd} and Sim_{ion}, we define a similarity function based on a re-interpretation of NGD. Let *a*, *b* be two variables (views, formal concepts or concepts). The network distance (ND) between *a* and *b* is given by:

$$\text{Network Dist}(a,b)=ND(a,b)=\frac{\max\{\log f(a),\log f(b)\}-\log f(a,b)}{\log(SizeNet)-\min\{\log f(a),\log f(b)\}} \tag{10}$$

The following definitions give the interpretations of *f*(*a*), *f*(*b*), *f*(*a*, *b*) and *SizeNet*. In this regard, it is worth mentioning that the probability of co-occurrence measured by the NDG when using the Web cannot be confounded with semantic similarity. For example, two words may co-occur very frequently in the web but not represent the same thing (e.g. geospatial and data). However, Sim-Net, while using the NDG formula, does not use the Web to assess co-occurrence of terms. Rather, it uses the NDG formula to measure the number of common and exclusive features of concepts with a formula that has proven to be adapted to networks.

Definition 1 (View Similarity Sim$_{\mathbf{view}}$) Consider two views *a* and *b*. *f*(*a*) is the number of views that are directly related to *a* in the network, plus the ones that are linked to them with generalisation relationships:

$$f(a) = \left|\left\{V_i \middle| (a \subseteq \exists R.V_i) \vee (V_i \subseteq V_j, a \subseteq \exists R.V_j) \vee (V_i \subseteq a)\right\}\right| \tag{11}$$

This set is interpreted as the number of occurrences of *a* because the definitions of all the views inside this set contain *a*. *f*(*a*, *b*) is the number of views that are related to *a* and *b*: $f(a,b) = f(a) \cap f(b)$. *SizeNet* is the total number of views in the network. We incorporate this distance in a semantic similarity measure which will compare all views of a pair of concepts C$_1$ and C$_2$. The distance can be interpreted as a semantic dissimilarity: it is zero when compared elements are the same, and increases when compared elements are different. According to the literature, a network distance can be transformed into a semantic similarity value using an exponentially decaying function (Schwering 2008); following this principle, Sim$_{\text{view}}$ (C$_1$, C$_2$) is given by

$$Sim_{view}(C_1, C_2) = \frac{1}{|C_1|}\sum_{i=1}^{|C_1|} \max{}_j\,(e^{-\lambda(a,b)\cdot ND(a,b)}) \quad \text{with} \quad 1 \le j \le |C_2| \tag{12}$$

with |C$_1$| the number of views of C$_1$, |C$_2$| the number of views of C$_2$. We introduce λ(*a*,*b*) as a factor that determine the importance of the pair *a*, *b* in the similarity measurement. Each concept has several views, and each view has a set of instances which is a subset of the concept's instances. We propose that the ratio of the number of instances of a view with respect to the number of instances of a concept is representative of the importance of this view compared to other views in the semantic similarity assessment. This principle is employed to compute λ(*a*,*b*):

$$\lambda(a,b)=\frac{\left|C_1^I\right|\cdot\left|C_2^I\right|}{\left|a^I\right|\cdot\left|b^I\right|} \tag{13}$$

This formula is simply a ratio of the number of instances of the concepts, and the number of instances of their views.

Definition 2 (Cross-domain Similarity Sim$_{\mathbf{cd}}$). Let *a* and *b* be formal concepts. Consider two concepts referenced to their respective formal concepts:

$$O_1 : C_1 \xrightarrow{\text{RefTo}} L : a \text{ and } O_2 : C_2 \xrightarrow{\text{RefTo}} L : b \text{ .}$$

f(*a*) is interpreted as the number of concepts of the network that are referenced to *a* and *SizeNet* the total number of concepts in the network. The more *a* and *b* are similar, the more C_1 and C_2 are similar:

$$Sim_{cd}(C_1,C_2)=e^{-ND(a,b)} \tag{14}$$

Definition 3 (Inter-ontology neighbourhood Similarity Sim$_{\mathbf{ion}}$). The neighbourhood of concepts across different ontologies can help to identify similar concepts of different domains. Consider a concept C_1 of O_1 (Figure 7.6). C_1 is linked to concepts of O_3 and O_4. The set of concepts from other ontologies to which C_1 is linked constitute the inter-ontology neighbourhood of C_1, denoted *ion*(C_1). We consider that C_1 is similar to C_9 if C_9 is similar to some concept of *ion*(C_1). Sim$_{\text{ion}}$ (C_1, C_2) is given by:

$$Sim_{ion}(C_1,C_2)=\begin{cases}1 & \text{if } \mathrm{C}_1 \equiv C_2\\ \dfrac{1}{\left|ion(C_2)\right|}\displaystyle\sum_{i=1}^{\left|ion(C_2)\right|} e^{\mathrm{ND}(\mathrm{C}_1,C_2)} & \text{otherwise}\end{cases} \tag{15}$$

with

$$C_j \in ion(C_2)$$

We take *a* and *b* to be concepts, and *f*(*a*) is the number of concepts of the network that directly related to *a*, plus their sub-concepts:

$$f(a) = \left|\{C_i \middle| (a \subseteq \exists R.C_i) \vee (C_i \subseteq C_j, a \subseteq \exists R.C_j) \vee (C_i \subseteq a)\}\right| \qquad (16)$$

SizeNet is the total number of concepts in the network.

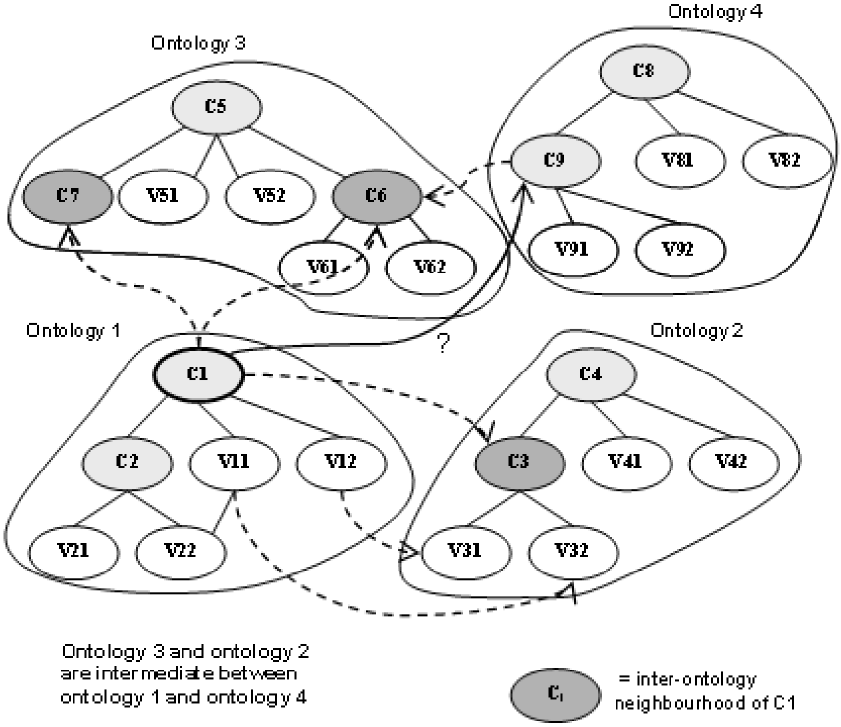

Ontology 3
Ontology 4
C5
C7
V51
V52
C6
V61
V62
C8
C9
V81
V82
V91
V92
Ontology 1
?
Ontology 2
C1
C4
C2
V11
V12
C3
V41
V42
V21
V22
V31
V32
Ontology 3 and ontology 2 are intermediate between ontology 1 and ontology 4
C_i = inter-ontology neighbourhood of C1

Figure 7.6 The inter-ontology neighbourhood of C_1

Some commonly discussed properties of similarity are the minimality (distance between from a concept to itself is zero), the symmetry (the distance from concept C1 to concept C2 is the same as vice-versa), and the triangle inequality (distance between two concepts C1 and C3 is always smaller than or equal to the distance between C1 and C2 plus the distance between C2 and C3). The NGD is symmetric, respects minimality but not triangle inequality (Cilibrasi and Vitanyi 2007). Sim-Net inherits the minimality and the non-respect of triangle inequality from the NGD. Sim-Net's respect of minimality is consistent with the fact that applying reasoning rules of Table 7.4 on two concepts having the same features will return the equivalence relation. Also, it cannot be assumed that the triangle inequality must be respected by a semantic distance.

However, Sim-Net is not symmetric, since in equations 10 and 13, we have introduced normalization factors that depend on the first concept C1 only (dividing by the number of views and the size of the neighbourhood of C1 respectively). This is consistent with the experiments of Tversky (1977) where it was shown that humans do not perceive similarity as symmetric. Also, it is consistent with the fact that semantic relations (equivalence, generalization, specialisation, etc) are not necessarily symmetric.

7.6.3.3.2 Computing Weights using Domain Similarity

The weights in Sim-Net semantic similarity measure reflect the similarity among domains. The domain is represented by the set of formal views and formal concepts to which a concept or view is referenced. We define the view domain, Dom(V), the concept domain, Dom(C), and the ontology domain, Dom(O). They are included in each other as follows:

$$Dom(V) \subseteq Dom(C) \subseteq Dom(O) \tag{17}$$

The ontology domain is the set of formal views and formal concepts to which all concepts and views of the ontology are referenced (Figure 7.7).

Definition 4 (Concept Domain *Dom*(*C*)). Consider the ontology sub-graph which starts from the root of the ontology, pass by the concept C_i and includes all sub-concepts of C_i:

$$\tau(O : C_i) = \{C, V | C \subseteq C_i \vee C_i \subseteq C, C \subseteq \exists HasView.V\} \tag{18}$$

Dom (C_i) is the set of all formal concepts and formal views to which the concepts and views included in τ(O: C_i) are referenced:

$$Dom(O_j : C_i) = \left\{ FC, FV \middle| \begin{matrix} C \xrightarrow{\text{ReferencedTo}} FC, C \in \tau(O:C_i), \\ V \xrightarrow{\text{ReferencedTo}} FV, V \in \tau(O:C_i) \end{matrix} \right\} \qquad (19)$$

This means that we consider that the domain is the set of entities that surround a concept in the ontology. To consider the domain of a concept will allow distinguishing between two concepts that would have, for instance, the same name and features, but that would be representing aspects of different realities, for example, a bridge as a road network element or a bridge as a hazard to boat navigation.

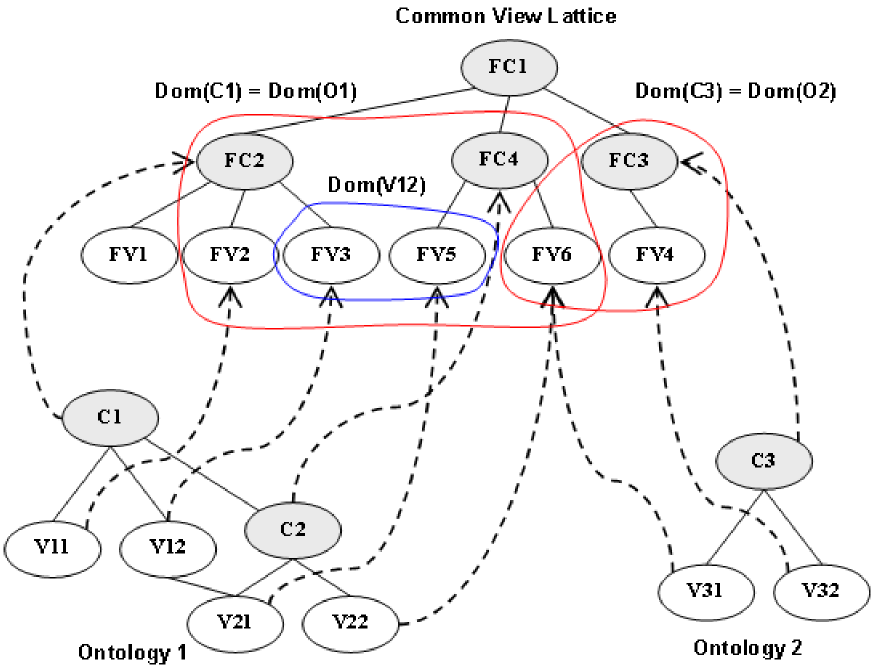


Figure 7.7 Ontology domain is a subset of the view lattice

Definition 5 (View Domain *Dom*(*V*)). *Dom*(V_i) is similar to *Dom* (C_i) but it contains only formal views:

$$Dom(O_j : V_i) = \left\{ FV \middle| V \xrightarrow{\text{ReferencedTo}} FV, V \in \tau(O:V_i) \right\} \qquad (20)$$

where $\tau(O:V_i)=\{V|V\subseteq V_i \vee V_i \subseteq V\}$. Figure 7.8 shows the sub-graphs $\tau(O:C_4)$ of a concept C_4 and $\tau(O:V_{42})$ of a view V_{42}. Their respective domains are defined by the set of formal views and formal concepts to which elements of their respective sub-graphs are referenced.

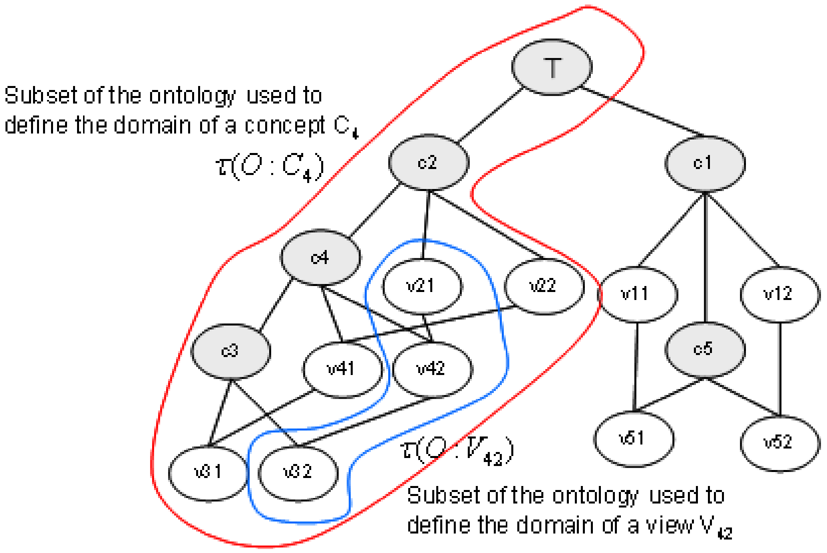

Subset of the ontology used to define the domain of a concept C_4
$\tau(O:C_4)$
T
c2
c1
c4
v21
v22
v11
v12
c3
v41
v42
c5
v31
v32
v51
v52
$\tau(O:V_{42})$
Subset of the ontology used to define the domain of a view V_{42}

Figure 7.8 Sub-graphs for the definition of Dom(C_4) and Dom(V_{42})

The domain similarity measure compares all formal views and formal concepts contained in the source and the target domains. Only pairs of formal concepts (or formal views) with maximal similarity are retained for the calculation of Sim_{dom}. The similarity between formal concepts FC_1 and FC_2 (or formal views) depends on the number of links that separates them from their most specific common subsumer MSCS. The latter is the most specific common parent of FC_1 and FC_2.

Definition 6 (Domain Similarity Sim_{dom}). To determine the similarity between two domains, we sum up the similarity of each pair of most similar formal concepts (or formal views) of the domain according to the following formula:

$$Sim_{dom}(Dom(C_1), Dom(C_2)) = \frac{1}{|Dom(C_1)|}\sum_i \max_j \left[\frac{2D_{ij}}{\delta_i + \delta_j + 2D_{ij}}\right]$$
$$\forall i, 1 \le i \le |Dom(C_1)| \qquad (21)$$

where *i* and *j* are subscripts for elements of Dom(C_1) and Dom(C_2) respectively, D_{ij} is the number of links from MSCS to the top of the view lattice, δ_i and δ_j are the number of links to MSCS. Sim_{dom} is used to compute weights in Sim-net, according to the next definition.

Definition 8 (Sim-Net Weights). The global semantic similarity expressed in equation 8 contains two weights: ω_{view} for similarity terms expected to be important when comparing concepts of similar domains, and ω_{cd} for the similarity terms that are expected to be important when comparing concepts of different domains. Therefore, we propose that ω_{view} should increase with domain while Sim_{dom} should decrease when domain similarity is low. The formulas for the weights express this:

$$\omega_{view}(C_1, C_2) = 1 - \left[\frac{1 - Sim_{dom}(C_1, C_2)}{\alpha}\right] \qquad (22)$$

$$\omega_{cd}(C_1, C_2) = \frac{1 - Sim_{dom}(C_1, C_2)}{\alpha} \qquad (23)$$

where α is a factor that can be used to reduce the importance of ω_{cd}. If we take α =1, we obtain that ω_{cd} = 1 when domains are completely dissimilar. In this case, it could be judged that we don't want to completely remove the influence of the sim_{view} term on the overall similarity. Therefore, we can increase the value of α. The weights computed with equations 22 and 23 are incorporated into global similarity equation (equation 8) to give the final similarity value. This is the result we get in a static network. In a dynamic, ad hoc, network, the adding of new ontologies may modify the similarities that were previously computed.

7.6.3.3.3 The Impact of "Ad Hoc" on the Sim-Net Similarity Measure

In this section, we examine the impact of a change in the ad hoc network on Sim-Net semantic similarity. When a change is detected, such as the adding of a source to the network or to a punctual cluster of ontologies, it is possible that existing semantic similarity values must be modified. This is not only because we have to compute the new similarity values between concepts of the new ontology with concepts of existing ontologies, but also because semantic similarity depends on inter-ontology neighbourhood. Consequently, the definition of Sim-Net itself is recursive in time: as the network evolves, previous values of similarity must be recomputed because the inter-ontology neighbourhood of concepts is modified. However, frequent changes in the network may require too much computation. As stated by Janowicz et al. (2008), semantic similarity measures are complex and most of the time costly in computation time, so approximation of similarity when changes occur is a most promising approach. We propose an algorithm that determines the behaviour of Sim-Net in the dynamic network. This algorithm is based on the following change-management criterion: when a new ontology O_{NEW} is added, we use the ontology domain similarity (Dom(O)) developed in section 4.3.3.2 to determine if the adding of O_{NEW} should modify the semantic similarity values that were computed between concepts of two other existing ontologies O_i and O_j. If Dom(O_{NEW}) is dissimilar to Dom(O_i), it is less probable that O_{NEW} will contain concepts that matches with those of O_i. Consequently, the chances that the inter-ontology neighbourhoods of O_i's concepts will be modified by the adding of O_{NEW} are low as well. We deduce that similarity values between O_i and any other ontology O_j would not be significantly modified by the adding of O_{NEW}. In this case, we decide not to re-compute those similarities. However, if Dom(O_{NEW}) and Dom(O_i) were enough similar (more than a given threshold Th_{DOM}), we expect that the adding of O_{NEW} will have a significant impact on the inter-ontology neighbourhoods of O_i's concepts, so we decide to re-compute similarity values between O_i and other ontology O_j, provided that Dom(O_j) is also similar to Dom(O_{NEW}).

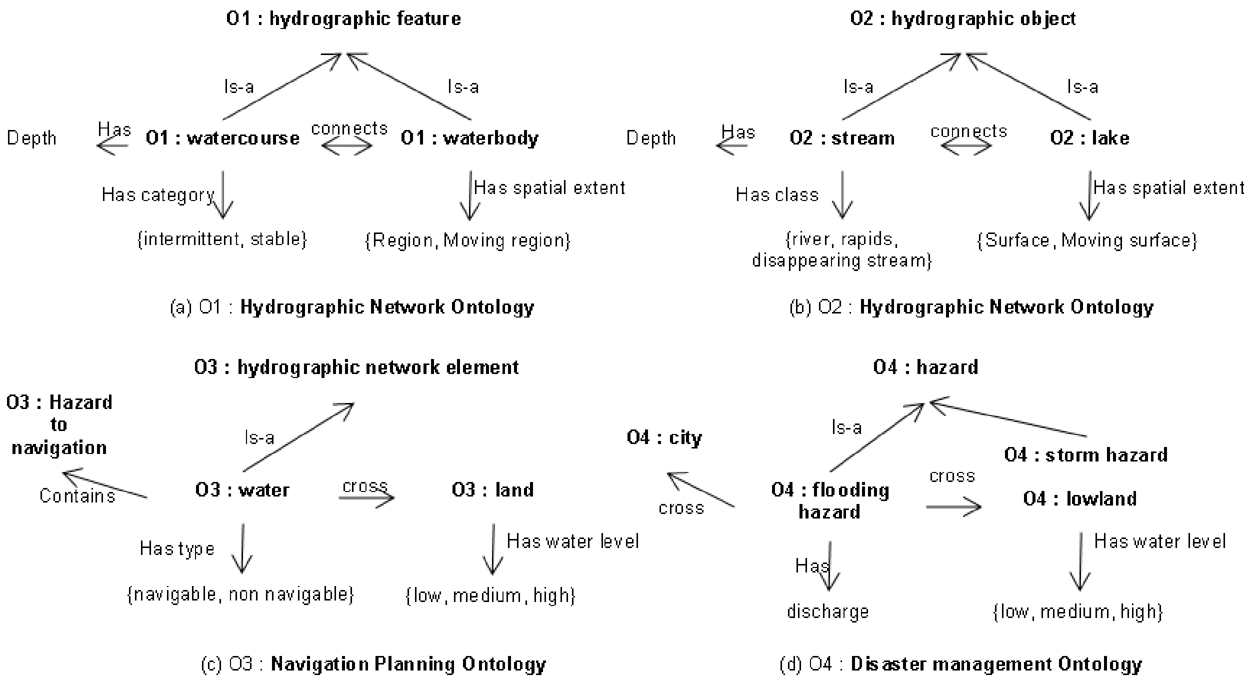
```
Dynamic Sim-Net Algorithm
Begin Algorithm
1  Detection of new node $O_{NEW}$ in the network or cluster
2  Regeneration of the View Lattice:
   2.1  Determine new reference concepts;
   2.2  Project local concept to new reference concepts;
   2.3  Identification of new formal concepts;
   2.4  Identification of new inheritance relationships and generation of the lattice;
   2.5  Determine new reference properties and values;
   2.6  Project local views to reference properties and values;
   2.7  Identification of new formal views;
   2.8  Identification of new inheritance relationships between formal views and expansion of the lattice
3 For each pair ($O_{NEW}$, $O_i$) of the network or cluster:
   3.1 Identify ontology domain Dom($O_i$) in view lattice;
   3.2 Compute ontology domain similarity $Sim_{DOM}(O_{NEW}, O_i)$;
   3.3 If $Sim_{DOM}(O_{NEW}, O_i) > Th_{DOM}$, compute semantic relationships between $O_{NEW}$ and $O_i$
4 For each $O_i$ of the network or cluster:
   4.1 For each $O_j$ , $j \neq i$, of the network or cluster:
      4.1.1 If concept(s) of $O_i$ and $O_j$ have non-disjoint semantic relations with $O_{NEW}$:
            4.1.1.1 Determine the new inter-ontology neighbourhood of those concepts;
            4.1.1.2 Re-compute the values in inter-ontology neighbourhood similarities, $sim_{ion}$
            4.1.1.3 Include new values of $sim_{ion}$ into global semantic similarity
End Algorithm
```

Figure 7.9 Dynamic Sim-Net Algorithm

The steps of the algorithm include also the regeneration of the view lattice, which is necessary in order to determine the new ontology domains (step 2). Step 3 verifies if ontology domains are similar enough while step 4 indicates the step for re-computing similarities. Note that we indicated that only the sim_{ion} term needs to be re-computed. In fact, the adding of new ontology may also have a slight impact on the values of the weights, since they depend on the domains, which may be slightly modified by the re-organization of the view lattice. The value of the chosen threshold Th_{DOM} is also a key factor that determines the efficiency of the algorithm. Experimental testing is required to fix the appropriated value.

7.7 Application Example

The following illustrates an example of the distinctive properties of Sim-Net. Consider the small ontologies in Figure 7.10. Ontologies O1 and O2 both describe hydrographic network. Ontologies O3 and O4 describe different domains. In the four ontologies we have concepts “watercourse”, “stream”, “water” and “flooding hazard” that represent different points of view on a watercourse.



(a) O1 : **Hydrographic Network Ontology**

(b) O2 : **Hydrographic Network Ontology**

(c) O3 : **Navigation Planning Ontology**

(d) O4 : **Disaster management Ontology**

Figure 7.10 Ontology examples

Watercourse and stream are constrained by the logic rules of Table 7.5, which allow extracting the views defined in second row. Table 7.6 gives the similarity of domain for the compared concepts, and Table 7.7 gives the semantic relationships and semantic similarity values obtained with Sim-Net.

Table 7.5 Logics rules and views for concepts of O1 and O2

Logic Rules	Extracted Views
O_1:	**O_1:**
r1: [depth(O1:watercourse) ≤ 20 m]	View1(O1:watercourse) :
→ [category(O1:watercourse) = intermittent]	category = intermittent
r2: [depth(O1:watercourse) > 20 m]	depth ≤ 20 m
→ [category(O1:watercourse) = stable]	spatial extent = moving region
r3: [category(O1:watercourse) = intermittent]	connect: O1:waterbody
→ [spatial extent(O1:watercourse) = moving region]	
r4: [category(O1:watercourse) = stable]	View2(O1:watercourse) :
→ [spatial extent(O1:watercourse) = region]	category = stable
	depth > 20 m
Inferred rules:	spatial extent = region
ir5: [depth(O1:watercourse) > 20 m]	connect: O1:waterbody
→ [spatial extent(O1:watercourse) = region]	
ir6: [depth(O1:watercourse) ≤ 20 m]	

→ [spatial extent(O1:watercourse)
= moving region]

O$_2$:
r1: [depth(O2: stream) ≤ 10 m] →
[class(O2: stream) = disappearing stream]
r2: [depth(O2: stream) > 10 m] →
[class(O2: stream) = river, rapids]
r3: [class(O2: stream) = disappearing stream] →
[spatial extent(O2: stream) = moving surface]
r4: [class(O2: stream) ≠ disappearing stream] →
[spatial extent(O2: stream) = surface]

Inferred rules:
ir5: [depth(O2: stream) > 10 m] →
[spatial extent(O2: stream) = surface]
ir6: [depth(O2: stream) ≤ 10 m] →
[spatial extent(O2: stream) = moving surface]

O$_2$:
View1(O2:stream) :
class = disappearing stream
depth ≤ 10 m
spatial extent = moving surface
connect: O2: lake

View2(O2:stream) :
class = river, rapids
depth > 10 m
spatial extent = surface
connect: O2: lake

By demonstrating the comparison of “O1: watercourse” and “O2: stream” we are illustrating the contribution of the view paradigm. View1(O1: watercourse) defines that an intermittent watercourse is a watercourse which depth is less than 20 meters and which spatial extent is a moving region, and in the same way view1(O2: stream) gives the semantic of “disappearing stream”. Only when these views are extracted it is possible to state that “O2: disappearing stream” is a kind of “O1: intermittent watercourse”, and doing the same for other views, find that “O2: stream” is more specific than “O1: watercourse”. Without extracting views, we find that “O1: watercourse” overlaps with “O2: stream” since we cannot compare “O1: intermittent watercourse” with “O2: disappearing stream”, and find them dissimilar. Without using views we have a lower semantic similarity value (0,48) than when using views (0,71). One important property of Sim-Net is its ability to consider spatial relationships and temporal relationships, in comparison to feature and geometric models. A feature model would find no similarity between ConnectWaterbody and ConnectLake since it allows no partial match. Sim-Net separates “connect” from “waterboby” and “lake” in the view lattice, finds that “lake” is a kind of “waterbody” and deduces that according to this property “O2: stream” is more specific than “O2: watercourse”.

Other examples show the specific properties of Sim-Net. The concepts “O1: watercourse” and “O3: water” do not have any common properties, however with cross-domain reasoning rules we find that “O1: watercourse” is a cross-domain specialisation of “O3: water” and their semantic similarity value is computed relatively to the domain similarity. Finally, “O1: watercourse” and “O4: flooding hazard” have no common properties and are not referenced to a common formal concept since they are lexically different. Common existing semantic similarity models would not find any similarity between these concepts. With similarity between inter-ontology neighbourhoods we find that “O4: flooding hazard” overlaps with one concept in the inter-ontology neighbourhood of “O1: watercourse”, i.e., “O3: water.” Therefore, Sim-Net detects semantic similarity between “O4: flooding hazard” and “O1: watercourse” with the help of a third ontology $O_{3,}$ which acts as an intermediary between O_1 and O_4. Furthermore, because domain similarity between “O4: flooding hazard” and “O1: watercourse” are dissimilar, high weight is given to the similarity between inter-ontology neighbourhoods.

Table 7.6 Similarity between concept domains of Figure 6.10

Concepts	Domain Similarity
Sim_{DOM} (O1:watercourse, O2: stream)	0,72
Sim_{DOM} (O1:watercourse, O3: water)	0,30
Sim_{DOM} (O3:water, O4: flooding hazard)	0,22
Sim_{DOM} (O1:watercourse, O4: flooding hazard)	0,0

Table 7.7 Similarity between concepts of Figure 7.10

Sim-Net semantic relationships	Semantic similarity value
$O_1 : \text{watercourse} \xrightarrow{\supset} O_2 : \text{stream}$	$SN = 0{,}71$ (with views) $SN = 0{,}48$ (without views)
$O_1 : \text{watercourse} \xrightarrow{\subset^*} O_3 : \text{water}$	$SN = 0{,}52$
$O_3 : \text{water} \xrightarrow{\cap} O_4 : \text{flooding hazard}$	$SN = 0{,}38$
$O_1 : \text{watercourse} \xrightarrow{\cap^*} O_4 : \text{flooding hazard}$	$SN = 0{,}42$

This demonstration shows some properties of Sim-Net in a static network. Now consider the adding of a new ontology O_{NEW} (Figure 7.11) to show the properties of Sim-Net in a dynamic network.

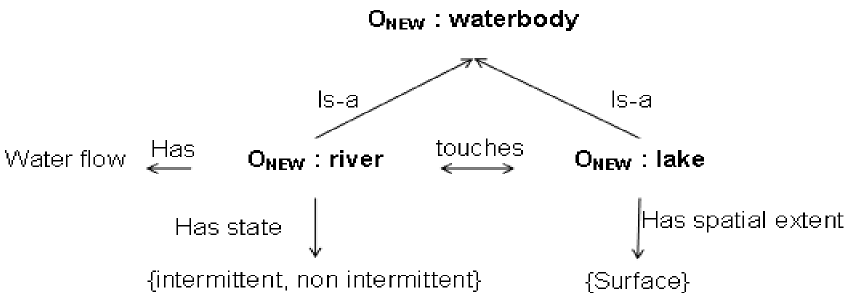


Figure 7.11 New ontology added to the network

The ontology domain similarities between O_{NEW} and the four other ontologies are shown in Table 7.8. We have to choose an appropriated threshold Th_{DOM}. For the purpose of the demonstration, we choose Th_{DOM} = 0,30 on the basis of previous results: concepts which domain similarity, according to Table 7.6, was at least 0,30 have a semantic similarity of more than 0,5 (Table 7.7). Note that this threshold is used only as an indication, and more investigation and extensive testing are required to choose an appropriate value. In this case, this would discard O3 and O4 from the re-computation of similarity values. Consequently, only the similarity between O1: watercourse and O2: stream is affected by the change. The new value of similarity (Table 7.9) increases compared to 0,72 because inter-ontology neighbourhood similarity was increased by the new ontology.

Table 7.8 Similarity between ontology domains

Ontologies	Domain Similarity
Sim_{DOM} (O1, O_{NEW})	0,65
Sim_{DOM} (O2, O_{NEW})	0,48
Sim_{DOM} (O3, O_{NEW})	0,20
Sim_{DOM} (O4, O_{NEW})	0,28

Table 7.9 New semantic similarity value

Concepts	Domain Similarity
Sim_{DOM} (O1: watercourse, O2: stream)	0,80

7.8 Conclusion and Future Work

While ad hoc networks of geospatial databases are becoming widespread, most existing semantic similarity models are targeted at comparing pairs of concepts in isolation. We have proposed a new semantic similarity model that contributes to two major issues. On the one hand, it addresses some requirements related to the assessment of semantic similarity in ad hoc network. In this regard, the new comparison criteria provided by Sim-Net include the comparison of inter-ontology neighbourhoods, and the comparison of domains. We also study the behaviour of Sim-Net when changes in the network occur, and propose that domain similarity be used as a criterion for deciding whether the adding of new ontologies will modify the existing similarity values. On the other hand, Sim-Net also copes with a rich representation of geospatial concepts were dependencies between properties can be represented and used to extract different views of a concept. The approach of assessing semantic similarity with Sim-Net is based on the establishment of a common semantic reference system built with the Formal Concept Analysis method. Sim-Net is also based on DL, which makes it readily adaptable to knowledge representation in existing ontologies as OWL is one of the most recommended ontology languages according to W3C.

Based on these findings, our research continues forwards as new research issues were raised during the development of Sim-Net. A first research issue that is specific to ad hoc networks is the propagation of error in semantic similarity assessment. This is likely to occur since in Sim-Net the similarity between two concepts depends on their similarity with neighbourhood concepts. In general, propagation of error can occur because concepts were ill-defined (low quality input), or because the semantic similarity measure employed is not suitable to the concept representation, or finally because results are inconsistent. This is an issue we have already explored in our previous work on elements of semantic quality (Bakillah et al. 2008). Therefore, how this framework can be incorporated with Sim-Net is a promising research avenue towards the resolution of the issue of error propagation. A second issue we planned to explore in future work is the problem of selecting which concepts in the network have to be compared. As it is not required that all pairs of concepts of several ontologies be compared to answer a given query, the question is how to propagate this query to relevant concepts and in the appropriate order. In this regard, we believe that Sim-Net could play a major role by indicating the relevant concepts. Finally, future work will include the comprehensive testing of Sim-Net in ad hoc networks.

CHAPTER 8

Implementation of the Approach with Real-Time-CPAR Prototype

8.1 Présentation du chapitre

Dans cette thèse, nous avons proposé un cadre conceptuel et une approche pour l'interoperabilité sémantique en temps réel dans un réseau ad hoc de bases de données géospatiales. Dans ce chapitre, nous présentons le prototype résultant de l'implantation de l'approche globale, ce qui constitue une première phase pour la validation de l'approche. Dans les sections suivantes, nous présentons l'architecture du prototype et les technologies utilisées. Puis, nous présentons les fonctionalités du prototype à l'aide d'un scénario qui démontre la validité et la faisabilité de l'approche.

8.2 Introduction

In this thesis, we have proposed a conceptual framework for real time semantic interoperability in ad hoc networks of geospatial databases. The contribution of this framework is to provide four main approaches, which are coalition discovering, query propagation, semantic augmentation, and semantic mapping, and to integrate these approaches to constitute the global real time semantic interoperability approach. In this chapter, we demonstrate the implementation of our approach with a prototype, called Real-Time-CPAR (Real-Time Coalition-Propagation-Augmentation-Reconciliation). In the next section, we present the technologies that were used to implement the prototype. Then, in section 8.4, we present the architecture of the prototype. Section 8.5 presents the interfaces of the prototype with an application example, in order to demonstrate the usefulness and contributions of the approach.

8.3 Technology Used for the Implementation

The proposed real time semantic interoperability approach is designed for a dynamic network environment where "nodes" can enter or quit the network, and form coalitions.

We have selected the JXTA platform as the core platform for implementing the framework[6]. JXTA simulates an open peer-to-peer network where independent peers can enter the network, send messages to each other and quit the network. The other characteristics of JXTA that explain this choice are the following:

- JXTA is developed in Java and is open source;
- It includes a basic, generic mechanism for the discovery of peers. This mechanism is used in the coalition discovering algorithm;
- It allows peers to form groups. This functionality was used in the coalition discovering algorithm, and the query propagation approach.
- The messages between peers, which are communicated through pipes, can carry different types of data (such as images, videos, queries, etc.).

Figure 8.1 shows the architecture of JXTA[7]. This architecture includes three layers: the core layer, the service layer and the application layer. The core layer encapsulates the building blocks that support peer-to-peer networking, including peer discovery, creation of peers, creation of peer groups, as well as security features. The service layer comprises several additional services, such as indexing and searching, storage systems, and file sharing. Third party services and applications are services and applications that can be added to the JXTA platform. The application layer implements applications including real time message passing between peers and sharing of resources.

[6] www.jxta.org

[7] Sun Microsystems. JXTA v2.3.x: Java^{TM} Programmer's Guide. JXTA project documentation, www.jxta.org, June 2010.

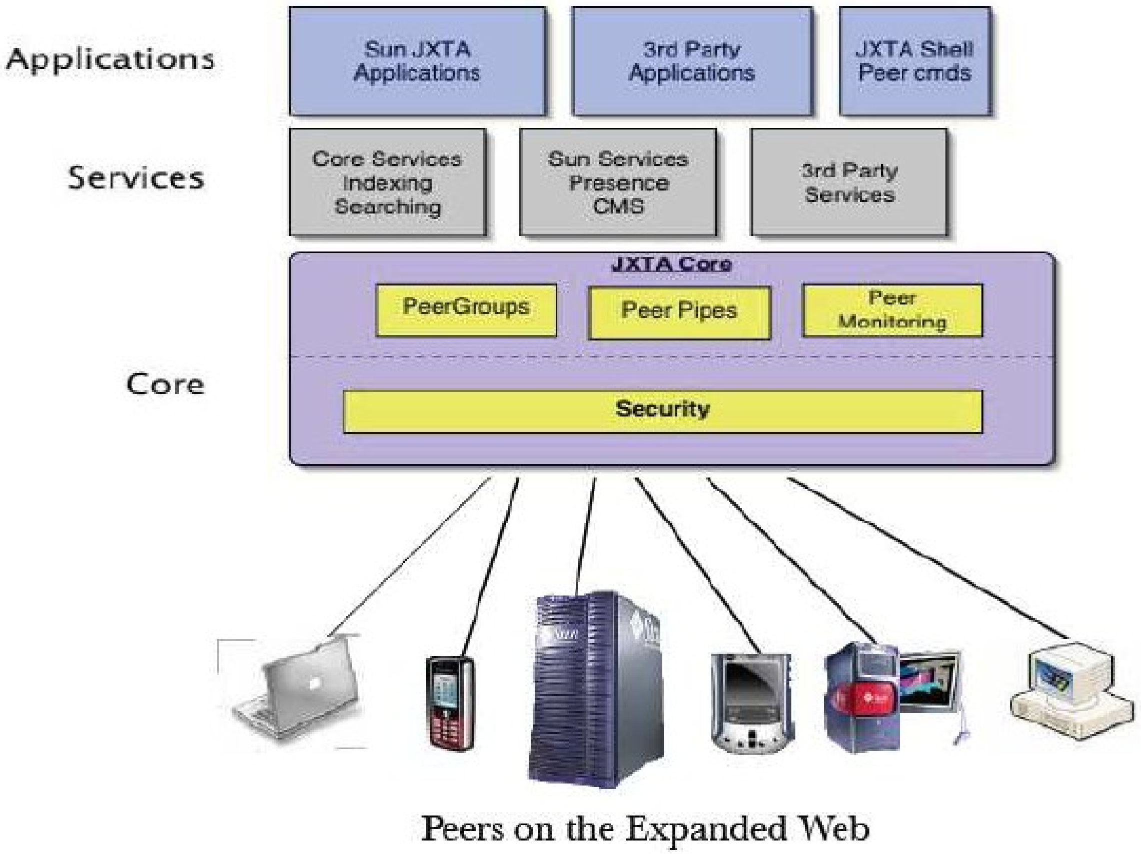


Figure 8.1 Architecture of JXTA, from Sun Microsystems

For a more comprehensive description of the JXTA platform, see Sun Microsystems, 2010.

8.4 Architecture

The architecture of the prototype is presented in Figure 8.2. The architecture is composed of four main components: the coalition discovery component, the real time query propagation component, the MVAC semantic augmentation component, and the semantic reconciliation component. Each peer is autonomous in term of knowledge representation, and processing capabilities. A local peer holds a geospatial database, and an ontology that specifies the semantics of data. The geospatial database can store data provided by real time, operational data sources, such as geo-sensor networks monitoring environmental conditions and mobile devices, and static data sources, such as static maps and repositories.

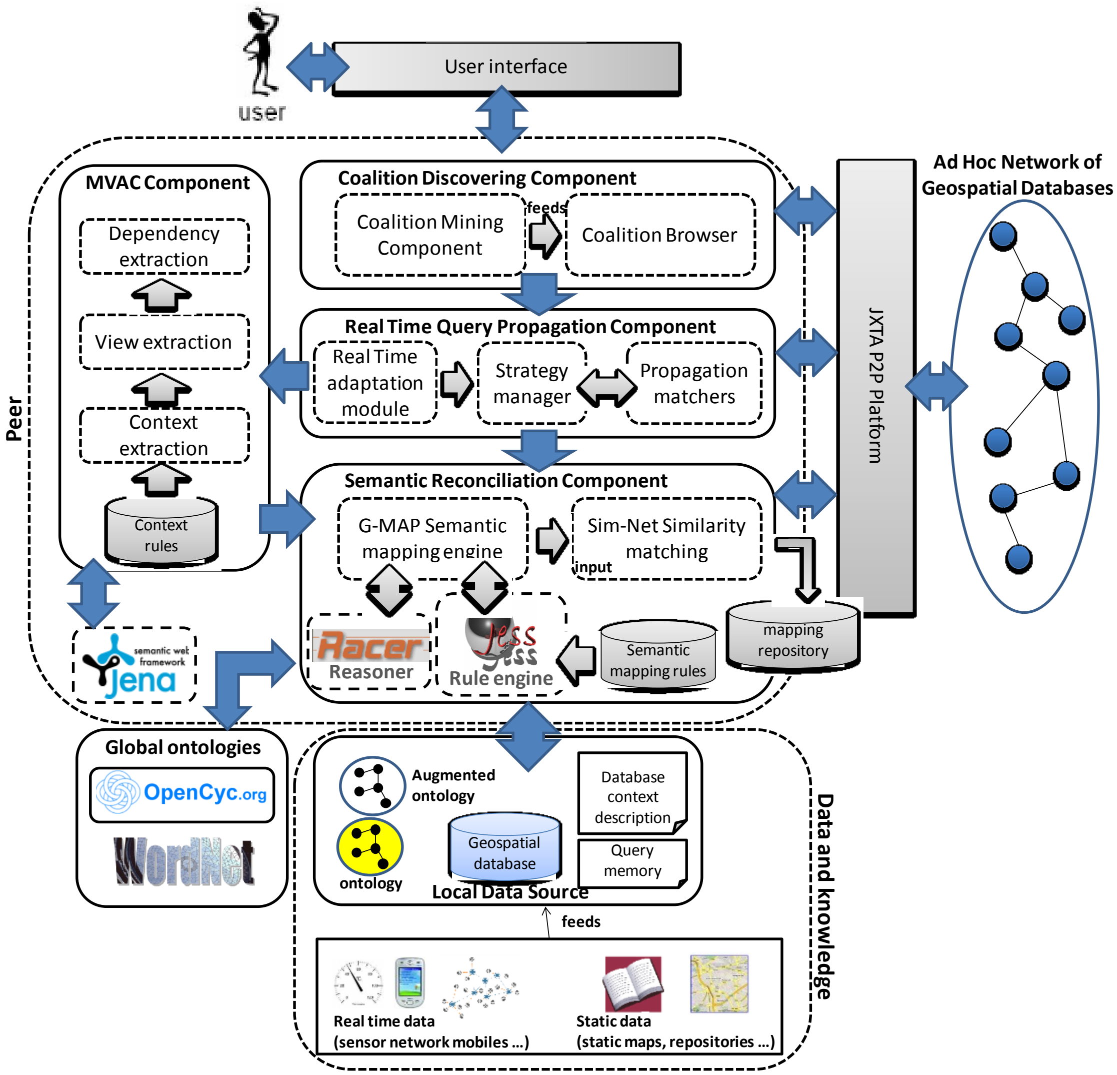


Figure 8.2 Architecture of the prototype

In addition, each peer holds the following information:

- an OWL description of the context of the database. The context is defined by (1) the geographical location covered by the entities stored in the database, (2) the period of time covered by those same entities, (3) the domain described by the database (ex: roads, water bodies, demography, etc.), (4) the role of the database (e.g., monitor water levels, characterize buildings, localise fire hazards, etc.).

- a list of queries it had previously answered. This list is called the peer's memory. It is used for query propagation. The memory of the peer is distinct from the context of the database.

The components are described as follows:

- The **coalition discovery component** implements the coalition mining algorithm described in Chapter 4. Its goal is to support the discovery of meaningful coalitions in the network. The group formation functionality of JXTA is used to materialize the coalitions. Through a coalition browser, the user can visualize the different coalitions available in the network, which are each described by a context. Moreover, the coalitions are used in the query propagation process.
- The **real time query propagation component** is responsible for forwarding users' queries to the relevant peers of the network. The user interface allows users to formulate queries. A query is composed of (1) the concept representing the data that the user is looking for, and (2) the context of the query, which allows to better identify the peers that can answer the query. The real time query propagation component implements the different propagation strategies described in Chapter 6, as well as the real time adaptation algorithm that reacts to changes in the network.
- The **MVAC semantic augmentation component** is responsible for enriching the ontologies of peers in order to support enhanced semantic mapping results and help the user to better understand the meaning of the retrieved data. It implements the semantic augmentation method described in Chapter 5. This component uses the Jena reasoner to support the extraction of dependencies. Jena is a tool that implements an instance-based reasoner for OWL ontologies. More specifically, it is used to verify the instances of concepts that verify dependencies.
- The **semantic reconciliation component** is responsible for finding semantic mappings and semantic similarities between concepts of peers' ontologies. It implements the G-MAP semantic mapping approach and the Sim-Net semantic similarity approach. G-MAP is a rule-based reasoning engine; the Jess rule-based reasoning engine is used to process the mapping rules and therefore infer the semantic relations between concepts.

The semantic reconciliation component also uses the Open Cyc spatial and temporal ontology, as well as the Word Net terminological database as global ontologies.

The four components are interrelated to support the global real time semantic interoperability process (RTSIP). The RTSIP starts with the computation of the coalitions. These coalitions are stored in order to be reused every time a query is issued. When a user submits a query, the query propagation component uses the coalitions to find the relevant peers that could answer the query. Once the query propagation graph is issued, the ontologies of the peers that are part of this propagation graph (and that were therefore selected as relevant peers) are sent to the MVAC component. The MVAC component semantically augments these ontologies. The augmented ontologies are sent to the semantic reconciliation component, which computes the semantic relations between the concepts of these ontologies. The user can use the semantic relations to choose the concept(s) that best represents the data he or she is looking for. The prototype was implemented in Java. The ontologies were developed in OWL with the Protégé ontology editor. This choice of language is motivated by the reasoning capabilities of OWL and by the fact that the OWL language is the W3C recommendation for Semantic Web applications.

8.5 Geospatial Data Used for the Implementation

For the implementation and testing of the prototype, we have chosen geospatial databases that were related to different aspects of disaster and risk management, including topographic databases, land management databases, flooding databases, etc. Disaster and risk management is a multi-disciplinary field that brings together experts and data from various fields. A massive amount of information relevant to disaster management is available, including cadastral information and administrative divisions, road networks and evacuation routes, hydrographic networks and flood zones, localization of environmentally sensitive facilities such as chemical industries, pipelines, and infected material disposal sites, health care facilities, weather data, and data on hazards such as seismic hazards and flooding hazards. These different data are often maintained by different organisms (government departments, non-governmental organizations (NGOs), academic research institute, private sector, industries, etc.) and for different purposes; they are therefore highly heterogeneous and conflicting. Identifying various perspectives and contexts, understanding the differences between information that conflicts or reconciling different representations for the same reality constitute a challenge for experts, who must make rapid, efficient and sound decisions during disasters in order to save lives and avoid material damage. To reflect this diversity of sources and perspectives, we have built database contexts and ontologies using selections of data products specifications of various geospatial data sources. Those sources include: the National Topographic Database of Canada (NTDB) (Natural Resources Canada, 1996), which contains information about roads, buildings, facilities, and waterbodies; the Quebec Topographic Database 1:20 000 (BDTQ) (Québec, 2000); diverse data sets on disasters risks and events (including flooding, earthquakes and tornados) in North America; the International Disaster Database (EM-DAT, 1988) and the Topographic and Administrative Database of Quebec (BDTA). To implement the prototype, we have developed a set of OWL geospatial database context descriptions, ontologies for databases, and instances, based on UML schemas and specifications of samples of the above data sets. Each class (concept) and relation of the data sources was documented with specifications, which were employed to build concept definitions. The spatial and temporal descriptors of concepts were defined manually based on textual definitions of concepts. Annex 3 provides an excerpt of the ontologies that we have developed.

8.6 Presentation of the Real-Time CPAR Prototype with a Case Study

In the following, we demonstrate the usefulness of the approach with a scenario where a user searches for *watercourses that are near a given residential area*, in order to *assess flooding risk in diverse regions of Canada in the last decade*. The final goal is to find the sources that can provide data to answer this query, and more specifically, the concepts of the database ontologies that are relevant to answer the query. The main obstacles to the achievement of this goal are the following:

- No global knowledge about the sources that are within the network and on the type of data they contain is available; so the user cannot know which source to query.
- The descriptions of sources at nodes are semantically heterogeneous, so finding the relevant sources that could contain data on watercourses or similar hydrographic features and factors of flooding risks is not straightforward.
- The ontologies of geospatial databases are heterogeneous, so different ontologies may represent concepts similar to flooding risks and watercourses in different ways; and semantics might be implicit. For example, the context may influence how a flooding risk region is spatially delineated. It is therefore difficult to retrieve the relevant concepts, but also to identify the heterogeneities that the user should be aware of in order to avoid misinterpretation of data.

In the following, it will be demonstrated how the prototype can help to overcome these obstacles to semantic interoperability. Figure 8.3 shows the main interface of the prototype. The four tabs which are accessible to any peer represent the four components of the global semantic interoperability: coalition discovery, query propagation, MVAC tool and G-MAP semantic reconciliation tool, which also integrate Sim-Net. At the beginning of the scenario, the user is facing a large number of sources. He or she does not know what kind of data they contain, so he or she cannot assess which source can provide the data he or she needs. As indicated above, the first step to resolve this problem is to find the coalitions in the ad hoc network. The coalitions partition the set of available sources into meaningful groups, and make the ad hoc network easier to search. In this step of the process, one of the difficulties to overcome is the heterogeneity between the descriptions of databases.

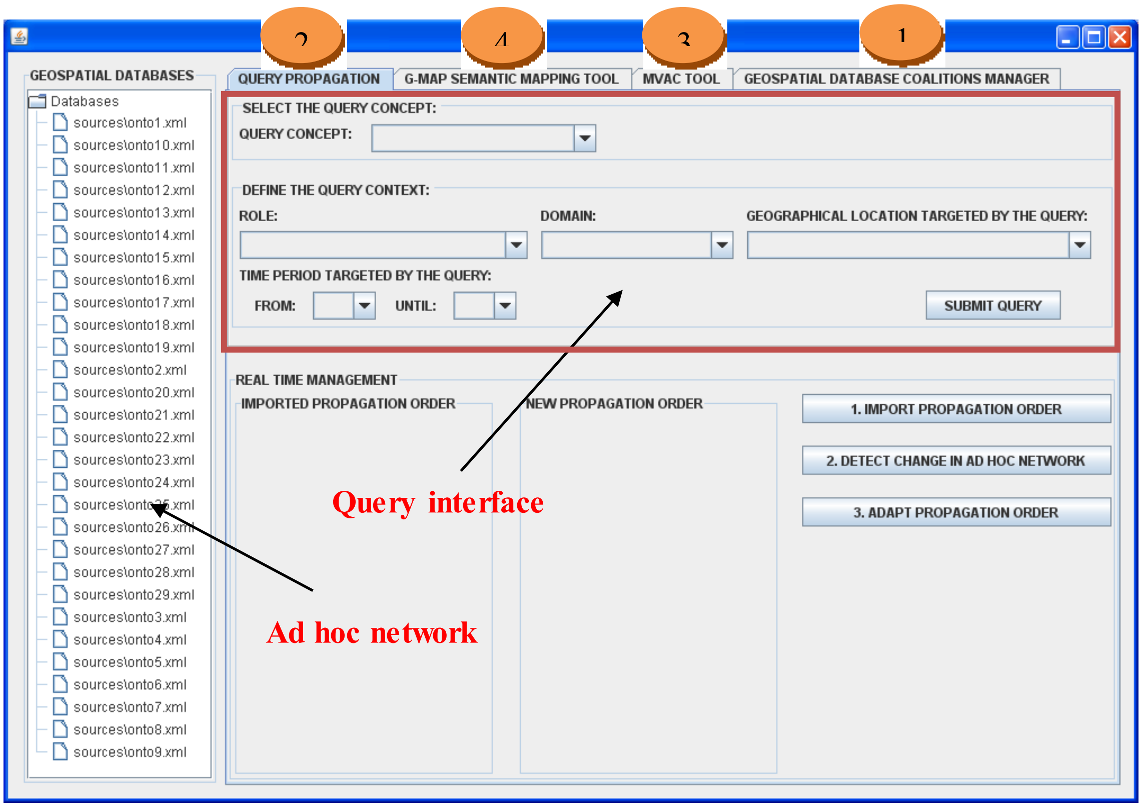


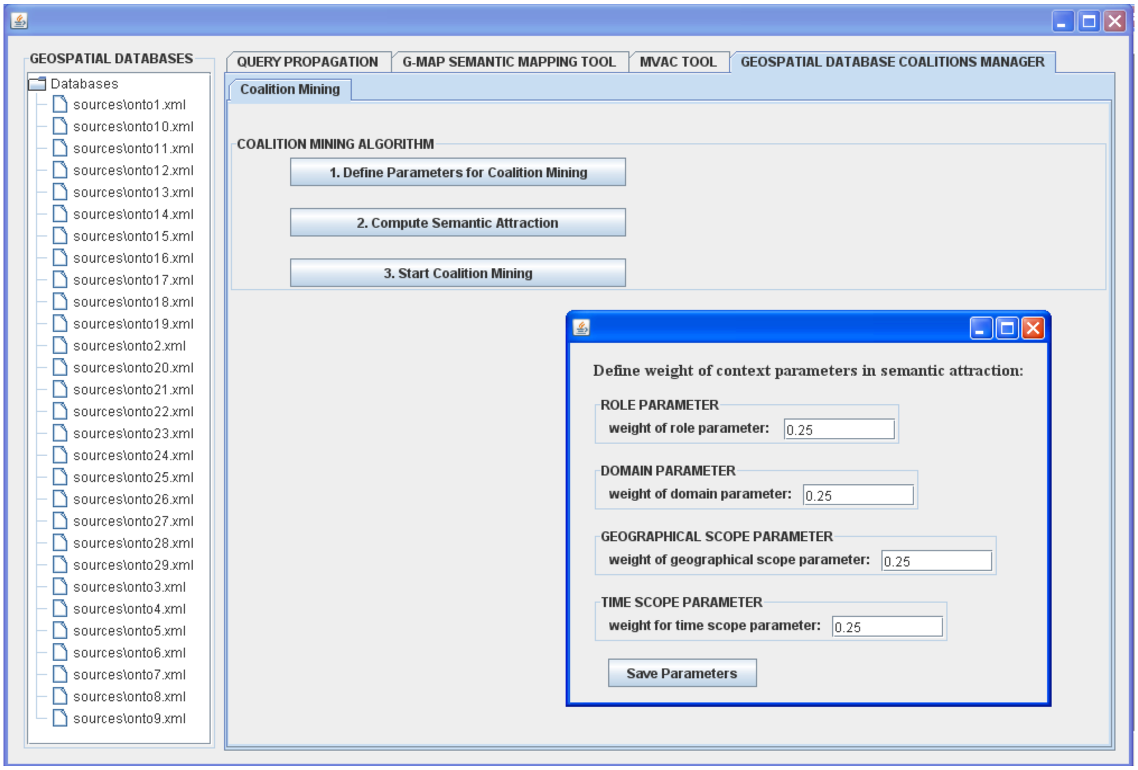
Figure 8.3 Main interface of the Real-Time CPAR prototype

8.6.1 Discovering Meaningful Coalitions

Figure 8.4 shows the main interface of the coalition discovery component.

The user goes through three main steps: (1) defining the parameters for coalition mining; (2) computing semantic attractions; and (3) running the coalition mining algorithm to find the coalitions.

In step (1), the user notably sets the weights that will be assigned to the parameters of the semantic attraction measure. A weight is given for each parameter of the database context. This allows the user to decide if one or some parameter(s) should be more decisive in the creation of the coalitions. For example, if the user wants to form coalitions only according to the function of the database (regardless of geographical location, domain, or temporality), he or she can set the weight for the function parameter to 1 and others to 0.



Define weight of context parameters in semantic attraction:

ROLE PARAMETER

weight of role parameter: 0.25

DOMAIN PARAMETER

weight of domain parameter: 0.25

GEOGRAPHICAL SCOPE PARAMETER

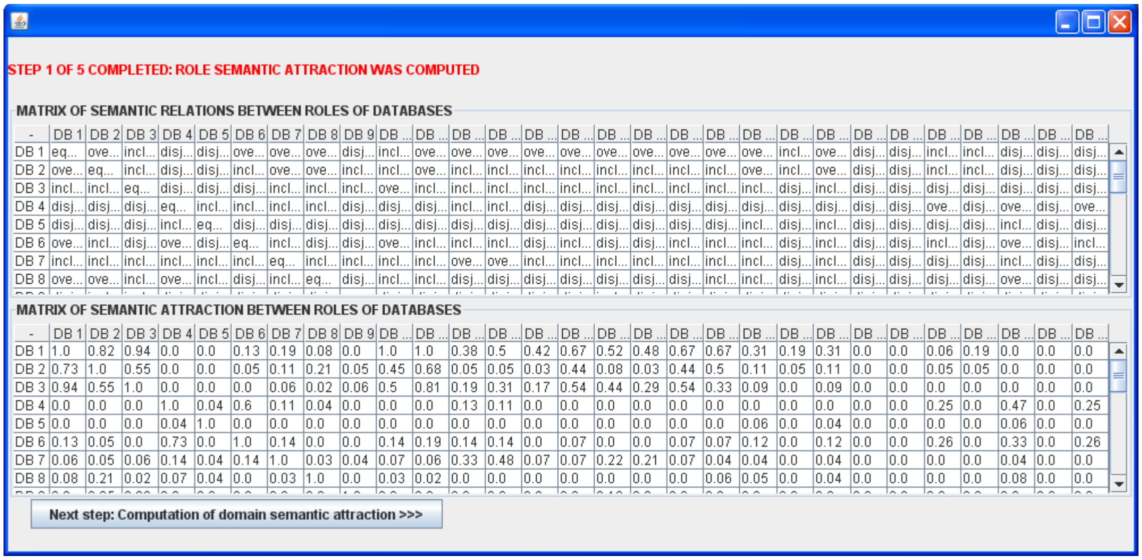
weight of geographical scope parameter: 0.25

TIME SCOPE PARAMETER

weight for time scope parameter: 0.25

Figure 8.4 Main interface of the coalition discovery component and definition of weights for computation of semantic attraction

In step (2), the user triggers the computation of semantic attraction between databases. The semantic attraction is computed according to the four context parameters, and summed to issue a matrix of global semantic attractions. For example, Figure 8.5 shows the matrix for the function-based semantic attraction. The computation of semantic attractions allows discovering the databases whose contexts are similar by resolving heterogeneities, including naming heterogeneities (e.g., inundation vs floods) and spatial granularity heterogeneities between context elements (e.g., Quebec included in Canada, road networks included in transportation network, and hydrography includes waterbodies). To build the global ontology of domains and functions, we exploited the Semantic Web for Earth and Environmental Terminology (SWEET) global ontologies, which are OWL ontologies developed by the NASA. Top-level concepts of SWEET include natural phenomena, processes, realms such as land surface, atmosphere, geosphere, hydrography, etc (which are useful to identify the domain), human activities (which are useful to identify functions), and qualities of substances, among others.



STEP 1 OF 5 COMPLETED: ROLE SEMANTIC ATTRACTION WAS COMPUTED

MATRIX OF SEMANTIC RELATIONS BETWEEN ROLES OF DATABASES

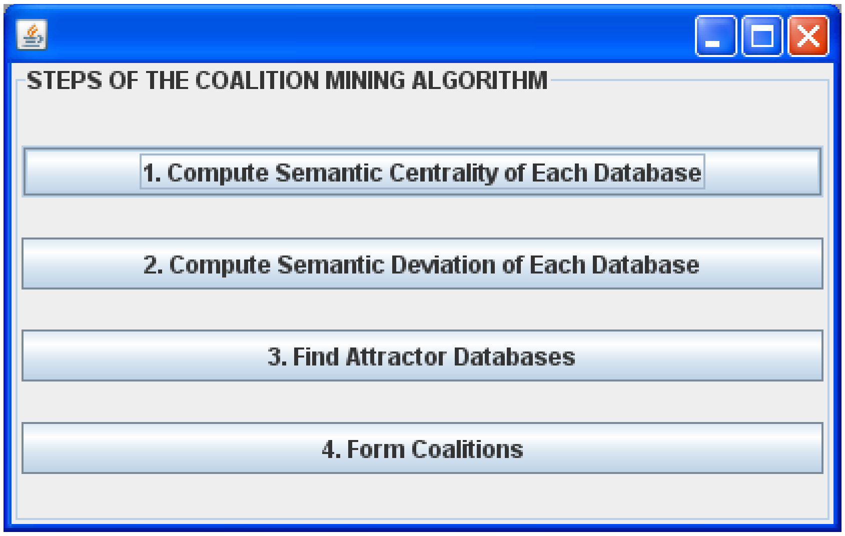
-	DB 1	DB 2	DB 3	DB 4	DB 5	DB 6	DB 7	DB 8	DB 9	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...
DB 1	eq...	ove...	incl...	disj...	disj...	ove...	ove...	ove...	disj...	incl...	ove...	ove...	ove...	ove...	ove...	ove...	ove...	ove...	ove...	ove...	incl...	ove...	disj...	disj...	incl...	incl...	disj...	disj...	disj...
DB 2	ove...	eq...	incl...	disj...	disj...	incl...	ove...	ove...	incl...	incl...	ove...	incl...	incl...	incl...	incl...	incl...	incl...	incl...	incl...	ove...	incl...	ove...	disj...	disj...	incl...	incl...	disj...	disj...	disj...
DB 3	incl...	incl...	eq...	disj...	disj...	disj...	incl...	incl...	incl...	ove...	incl...	incl...	incl...	incl...	incl...	incl...	incl...	incl...	incl...	incl...	disj...	incl...	disj...	disj...	disj...	disj...	disj...	disj...	disj...
DB 4	disj...	disj...	disj...	eq...	incl...	incl...	incl...	incl...	disj...	disj...	disj...	incl...	incl...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	ove...	disj...	ove...	disj...	ove...
DB 5	disj...	disj...	disj...	incl...	eq...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	incl...	disj...	incl...	disj...	disj...	disj...	disj...	incl...	disj...	disj...
DB 6	ove...	incl...	disj...	ove...	disj...	eq...	incl...	disj...	disj...	ove...	incl...	incl...	incl...	disj...	incl...	disj...	disj...	incl...	incl...	incl...	disj...	incl...	disj...	disj...	incl...	disj...	ove...	disj...	incl...
DB 7	incl...	incl...	incl...	incl...	incl...	incl...	eq...	incl...	incl...	incl...	incl...	ove...	ove...	incl...	incl...	incl...	incl...	incl...	incl...	incl...	disj...	incl...	disj...	disj...	disj...	disj...	incl...	disj...	disj...
DB 8	ove...	ove...	incl...	ove...	incl...	disj...	incl...	eq...	disj...	incl...	incl...	disj...	disj...	disj...	disj...	disj...	disj...	disj...	incl...	incl...	disj...	incl...	disj...	disj...	disj...	disj...	ove...	disj...	disj...

MATRIX OF SEMANTIC ATTRACTION BETWEEN ROLES OF DATABASES

-	DB 1	DB 2	DB 3	DB 4	DB 5	DB 6	DB 7	DB 8	DB 9	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...	DB ...
DB 1	1.0	0.82	0.94	0.0	0.0	0.13	0.19	0.08	0.0	1.0	1.0	0.38	0.5	0.42	0.67	0.52	0.48	0.67	0.67	0.31	0.19	0.31	0.0	0.0	0.06	0.19	0.0	0.0	0.0
DB 2	0.73	1.0	0.55	0.0	0.0	0.05	0.11	0.21	0.05	0.45	0.68	0.05	0.05	0.03	0.44	0.08	0.03	0.44	0.5	0.11	0.05	0.11	0.0	0.0	0.05	0.05	0.0	0.0	0.0
DB 3	0.94	0.55	1.0	0.0	0.0	0.0	0.06	0.02	0.06	0.5	0.81	0.19	0.31	0.17	0.54	0.44	0.29	0.54	0.33	0.09	0.0	0.09	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DB 4	0.0	0.0	0.0	1.0	0.04	0.6	0.11	0.04	0.0	0.0	0.0	0.13	0.11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.25	0.0	0.47	0.0	0.25
DB 5	0.0	0.0	0.0	0.04	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.06	0.0	0.04	0.0	0.0	0.0	0.0	0.06	0.0	0.0
DB 6	0.13	0.05	0.0	0.73	0.0	1.0	0.14	0.0	0.0	0.14	0.19	0.14	0.14	0.0	0.07	0.0	0.0	0.07	0.07	0.12	0.0	0.12	0.0	0.0	0.26	0.0	0.33	0.0	0.26
DB 7	0.06	0.05	0.06	0.14	0.04	0.14	1.0	0.03	0.04	0.07	0.06	0.33	0.48	0.07	0.07	0.22	0.21	0.07	0.04	0.04	0.0	0.04	0.0	0.0	0.0	0.0	0.04	0.0	0.0
DB 8	0.08	0.21	0.02	0.07	0.04	0.0	0.03	1.0	0.0	0.03	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.06	0.05	0.0	0.04	0.0	0.0	0.0	0.0	0.08	0.0	0.0

Figure 8.5 Semantic attraction matrix with respect to the function of databases

In step (3), the user is invited to follow several steps of the coalition mining process (Figure 8.6). The first three steps lead to the identification of the attractor databases.

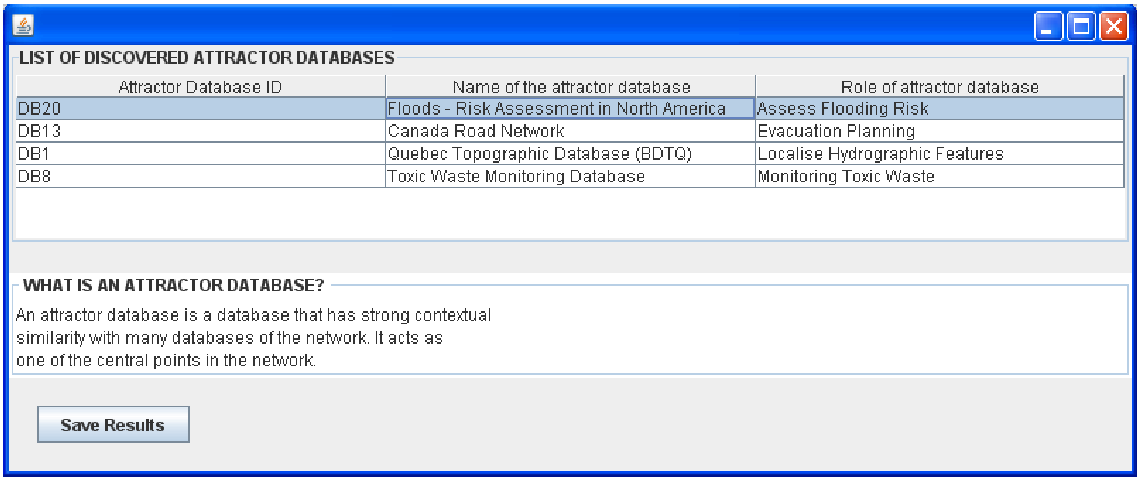


STEPS OF THE COALITION MINING ALGORITHM

1. Compute Semantic Centrality of Each Database
2. Compute Semantic Deviation of Each Database
3. Find Attractor Databases
4. Form Coalitions

Figure 8.6 The user follows the steps of the coalition mining algorithm

In this case study, four attractor databases were identified. They were identified as attractor databases because their respective context encompasses the contexts of several other databases. For example, with the global ontology on domains and functions, the semantic attraction is able to identify that the function “flood risk assessment” includes the function “to localize waterbodies” and therefore to include the databases that have this function in the coalition for “flood risk assessment.” The attractor databases are displayed to the user, who can vizualise the elements of their context, such as the function parameter (Figure 8.7).



LIST OF DISCOVERED ATTRACTOR DATABASES

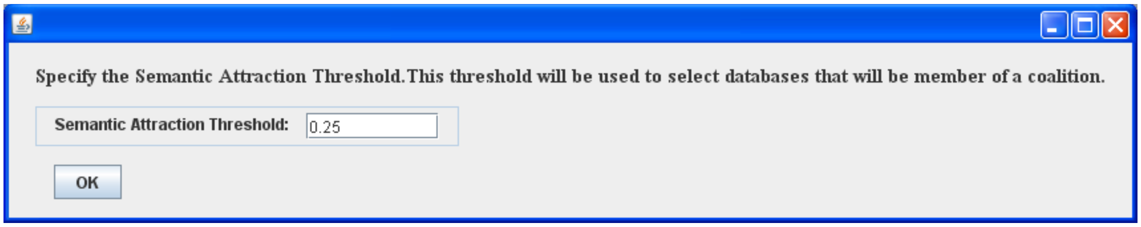
Attractor Database ID	Name of the attractor database	Role of attractor database
DB20	Floods - Risk Assessment in North America	Assess Flooding Risk
DB13	Canada Road Network	Evacuation Planning
DB1	Quebec Topographic Database (BDTQ)	Localise Hydrographic Features
DB8	Toxic Waste Monitoring Database	Monitoring Toxic Waste

WHAT IS AN ATTRACTOR DATABASE?

An attractor database is a database that has strong contextual
similarity with many databases of the network. It acts as
one of the central points in the network.

Figure 8.7 The user vizualises the attractor databases that were identified

The fourth step of the coalition mining algorithm is the formation of coalitions. For each attractor database, a coalition is formed with the peers whose context is included within the context of the attractor database. The user is invited to set a semantic attraction threshold above which other peers can be part of a coalition (Figure 8.8).

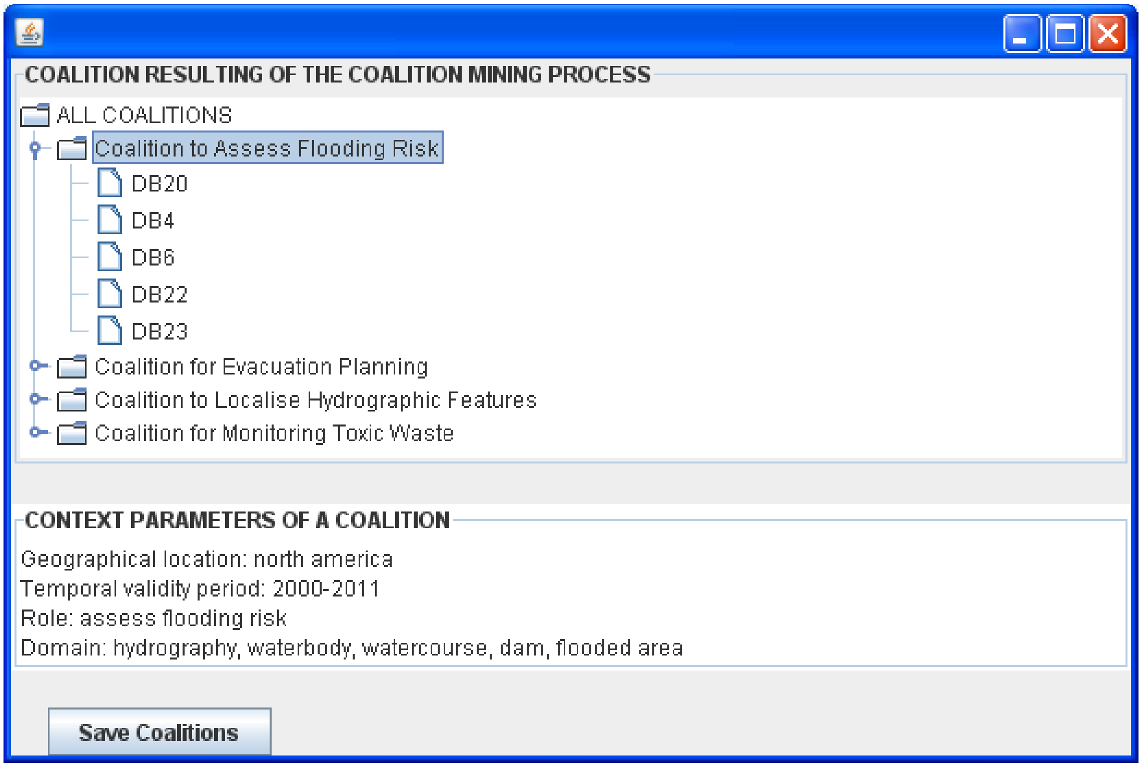


Specify the Semantic Attraction Threshold.This threshold will be used to select databases that will be member of a coalition.

Semantic Attraction Threshold: 0.25

Figure 8.8 The user sets the semantic attraction threshold to select the peers that can be part of a coalition

The advantage of the user-defined threshold is that the user can decide if he or she wants very "exclusive" coalitions, where members must meet very strict similarity conditions, or conversely, "loose" coalitions, which are more inclusive. The consequence is that the search for the relevant source is either narrowed or broadened. As a result, we may also obtain overlapping coalitions, i.e., coalitions which have common members. The lower the semantic attraction threshold, the more it is likely that resulting coalitions will overlap. Figure 8.9 shows the coalitions that were formed in the case study.



COALITION RESULTING OF THE COALITION MINING PROCESS

- ALL COALITIONS
  - Coalition to Assess Flooding Risk
    - DB20
    - DB4
    - DB6
    - DB22
    - DB23
  - Coalition for Evacuation Planning
  - Coalition to Localise Hydrographic Features
  - Coalition for Monitoring Toxic Waste

CONTEXT PARAMETERS OF A COALITION

Geographical location: north america
Temporal validity period: 2000-2011
Role: assess flooding risk
Domain: hydrography, waterbody, watercourse, dam, flooded area

Figure 8.9 The user visualizes the discovered coalitions, their members and their context

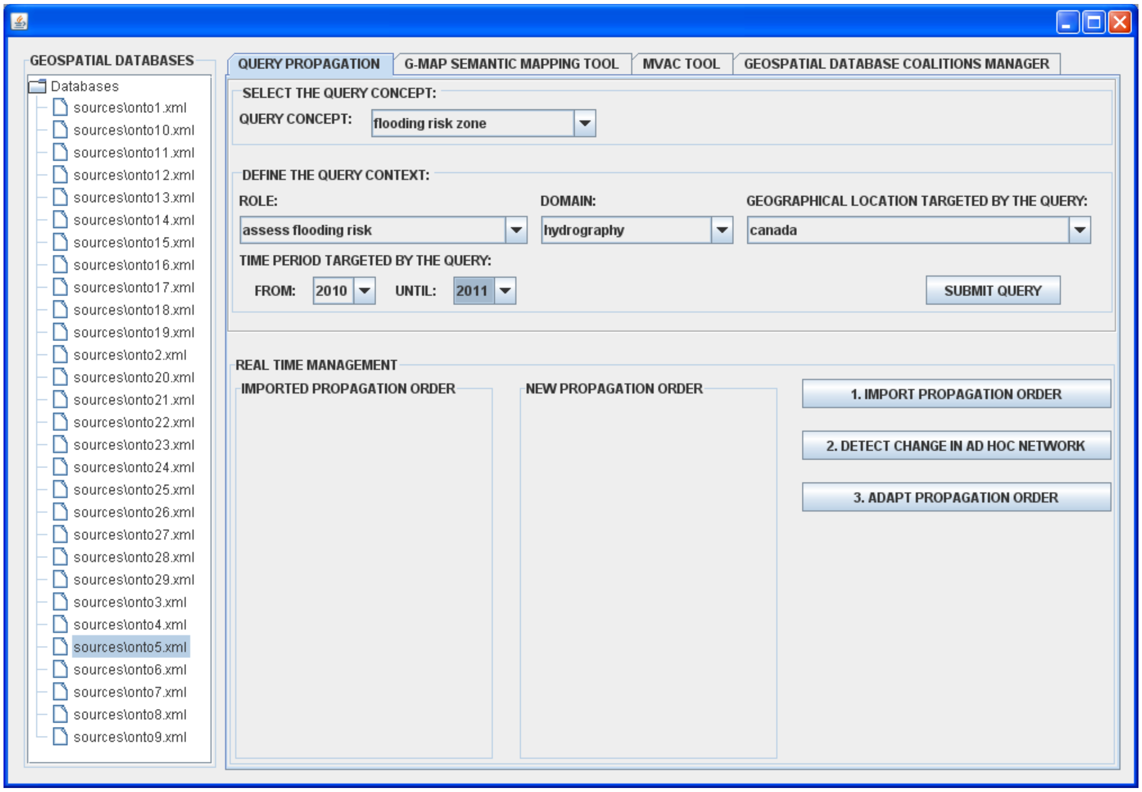
The coalitions that were formed help the user to understand, at a glance, the nature of data available in the network. It is a first step toward identifying the databases that are able to answer the query. However, not all those coalitions or the sources that they contain are relevant to the user's query. To avoid letting the user assess each source they contain, in the next step, the user will be invited to formulate his or her query and propagate the query in the network to find the relevant sources.

8.6.2 Real Time Query Propagation

In a decentralized ad hoc network, we cannot rely, for example, on a centralized repository of source metadata that could be queried to retrieve relevant sources. The query has to be propagated through the network. To do so, the problem is how to use the structure of the network (coalitions) and the knowledge on each node in an efficient manner, while resolving the heterogeneities between the user's query and the knowledge we rely on for query propagation. One of the distinctions of our approach with respect to other query propagation approaches (e.g., Zhuge et al. 2004; Montanelli and Castano 2008) is that it does not rely on existing semantic mappings between concepts of database ontologies. Instead, it relies on the context of coalitions and the contexts of sources. Only when peers to which the query should be forwarded are selected, this selection is further refined by computing semantic mappings between the query concept and the concepts of the selected peers' ontologies. The resulting advantage is that a smaller number of semantic mappings are computed to find the relevant data, reducing the cost of the semantic mapping process.

First, the user formulates a query. The query is composed of a concept and of the context of the query. For example, the following query is about the concept "flooding risk zone," in the geographical location "Canada," during the time period 2010-2011, with "hydrography" as domain and "produce risk map" as a function (in OWL RDF/XML syntax):

```
<owl:Class rdf:ID="FloodingRiskZone"/>
<owlx:ObjectProperty owlx:name="InGeographicalLocation">
      <owlx:domain owlx:class="FloodingRiskZone"/>
      <owlx:range owlx:class="Canada"/>
</owlx:ObjectProperty>
<owlx:ObjectProperty owlx:name="TimePeriodBegin">
      <owlx:domain owlx:class="FloodingRiskZone"/>
      <owlx:range owlx:class="2010"/>
</owlx:ObjectProperty>
<owlx:ObjectProperty owlx:name="TimePeriodEnd">
      <owlx:domain owlx:class="FloodingRiskZone"/>
      <owlx:range owlx:class="2011"/>
</owlx:ObjectProperty>
<owlx:ObjectProperty owlx:name="InDomain">
      <owlx:domain owlx:class="FloodingRiskZone"/>
      <owlx:range owlx:class="Hydrography"/>
</owlx:ObjectProperty>
<owlx:ObjectProperty owlx:name="Function">
      <owlx:domain owlx:class="FloodingRiskZone"/>
      <owlx:range owlx:class="ProduceRiskMap"/>
</owlx:ObjectProperty>
```



GEOSPATIAL DATABASES
QUERY PROPAGATION
G-MAP SEMANTIC MAPPING TOOL
MVAC TOOL
GEOSPATIAL DATABASE COALITIONS MANAGER
SUBMIT QUERY
REAL TIME MANAGEMENT
IMPORTED PROPAGATION ORDER
NEW PROPAGATION ORDER
1. IMPORT PROPAGATION ORDER
2. DETECT CHANGE IN AD HOC NETWORK
3. ADAPT PROPAGATION ORDER

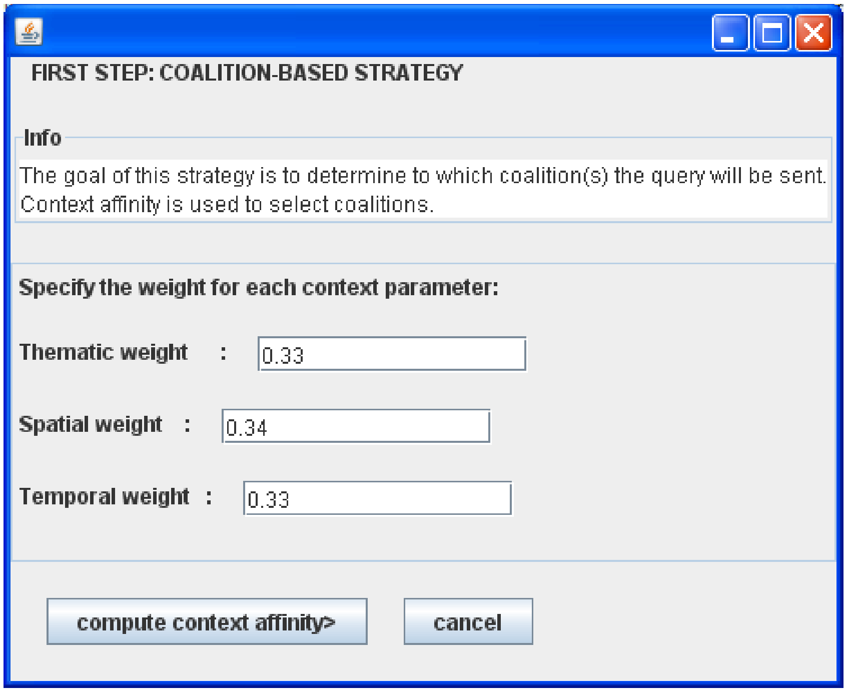
Databases
- sources\onto1.xml
- sources\onto10.xml
- sources\onto11.xml
- sources\onto12.xml
- sources\onto13.xml
- sources\onto14.xml
- sources\onto15.xml
- sources\onto16.xml
- sources\onto17.xml
- sources\onto18.xml
- sources\onto19.xml
- sources\onto2.xml
- sources\onto20.xml
- sources\onto21.xml
- sources\onto22.xml
- sources\onto23.xml
- sources\onto24.xml
- sources\onto25.xml
- sources\onto26.xml
- sources\onto27.xml
- sources\onto28.xml
- sources\onto29.xml
- sources\onto3.xml
- sources\onto4.xml
- sources\onto5.xml
- sources\onto6.xml
- sources\onto7.xml
- sources\onto8.xml
- sources\onto9.xml

SELECT THE QUERY CONCEPT:
QUERY CONCEPT: flooding risk zone

DEFINE THE QUERY CONTEXT:
ROLE: assess flooding risk
DOMAIN: hydrography
GEOGRAPHICAL LOCATION TARGETED BY THE QUERY: canada
TIME PERIOD TARGETED BY THE QUERY:
FROM: 2010 UNTIL: 2011

Figure 8.10 The user formulates the query for propagation

The first step in query propagation is to find the coalition(s) to which the query will be sent. This step is based on context affinity between the query and the coalitions' context, which will resolve semantic heterogeneities between the query's element and the description of coalitions' context. However, the user might find more important, for example, to consider the function and domain of the coalitions when selecting which ones are relevant. Therefore, the user can define weights which determine the importance of the thematic, spatial and temporal context parameters when comparing contexts (Figure 8.11). In the case study, the coalition to assess flooding risk is selected (Figure 8.12). However, the user can choose the threshold above which coalitions will be selected. The user can visualize the spatial, temporal and thematic affinities to support his or her decision when setting the threshold.



FIRST STEP: COALITION-BASED STRATEGY

Info

The goal of this strategy is to determine to which coalition(s) the query will be sent. Context affinity is used to select coalitions.

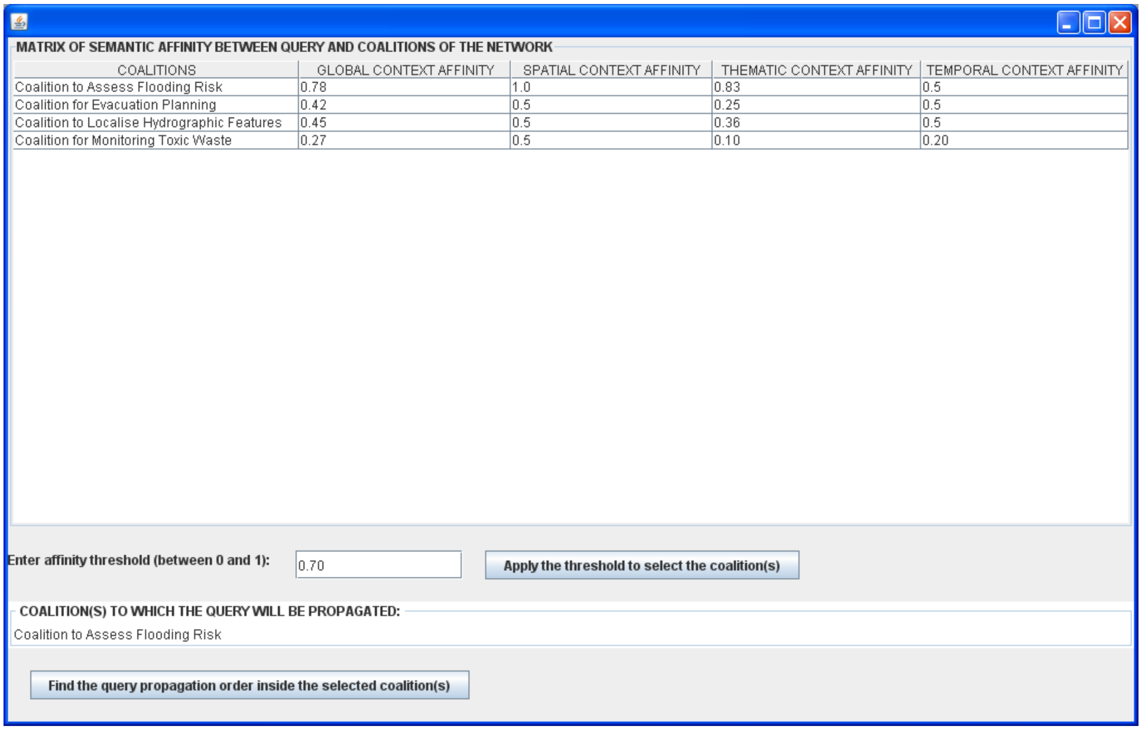
Specify the weight for each context parameter:

Thematic weight : 0.33

Spatial weight : 0.34

Temporal weight : 0.33

Figure 8.11 The user sets the weights for context affinity



MATRIX OF SEMANTIC AFFINITY BETWEEN QUERY AND COALITIONS OF THE NETWORK

COALITIONS	GLOBAL CONTEXT AFFINITY	SPATIAL CONTEXT AFFINITY	THEMATIC CONTEXT AFFINITY	TEMPORAL CONTEXT AFFINITY
Coalition to Assess Flooding Risk	0.78	1.0	0.83	0.5
Coalition for Evacuation Planning	0.42	0.5	0.25	0.5
Coalition to Localise Hydrographic Features	0.45	0.5	0.36	0.5
Coalition for Monitoring Toxic Waste	0.27	0.5	0.10	0.20

Enter affinity threshold (between 0 and 1): 0.70

COALITION(S) TO WHICH THE QUERY WILL BE PROPAGATED:

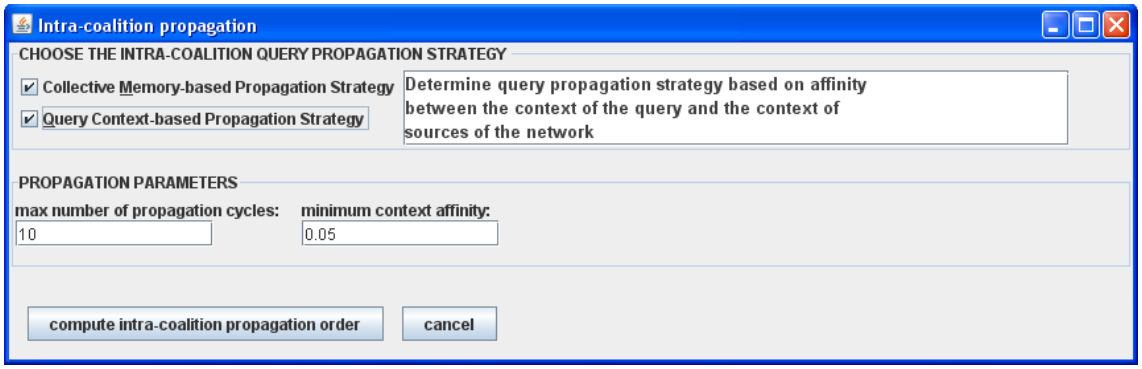
Coalition to Assess Flooding Risk

Find the query propagation order inside the selected coalition(s)

Figure 8.12 Affinities used to select the relevant coalition

However, not all the peers within the coalition are necessarily relevant to the query. Therefore, we need to refine the selection by performing propagation within the coalition.

Propagation within the coalition is based on two strategies: context or memory-based. The first is based on context affinity between the query and peers. The second is based on affinity between the current query and the past queries that were answered by peers. The user is invited to select one of the strategies, or a combination of both to maximize the result (Figure 8.13).


Intra-coalition propagation
compute intra-coalition propagation order
cancel

CHOOSE THE INTRA-COALITION QUERY PROPAGATION STRATEGY

☑ Collective Memory-based Propagation Strategy

☑ Query Context-based Propagation Strategy

Determine query propagation strategy based on affinity between the context of the query and the context of sources of the network

PROPAGATION PARAMETERS

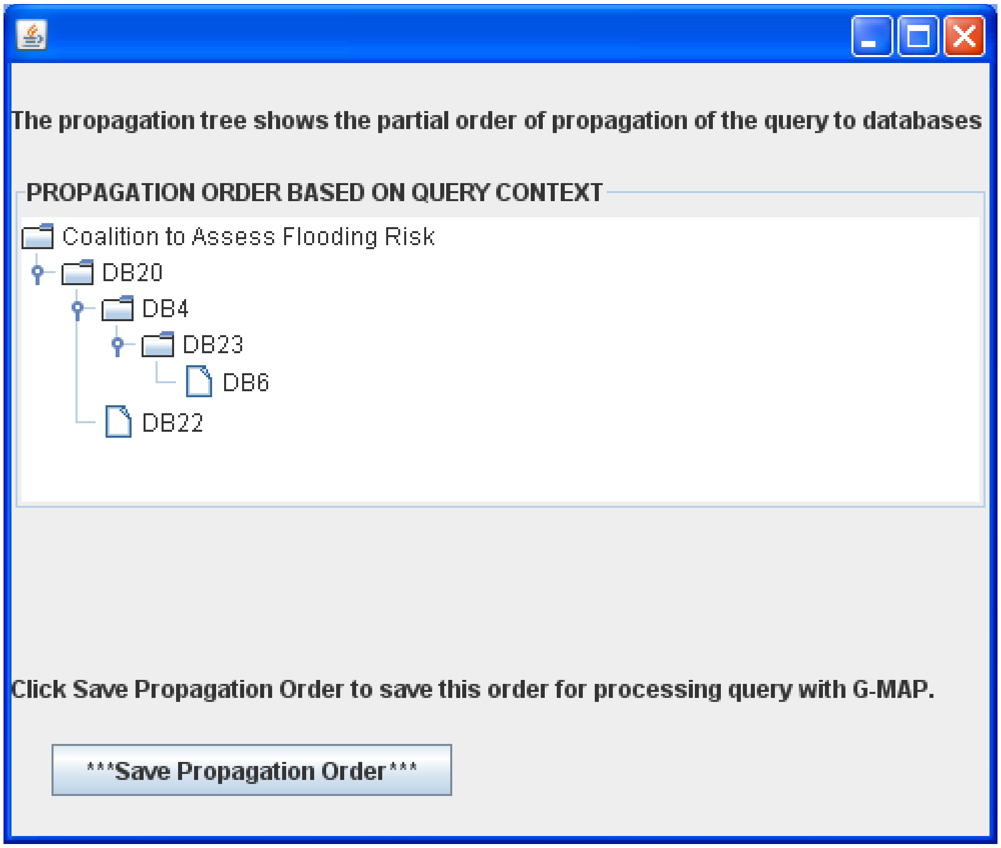
max number of propagation cycles: 10

minimum context affinity: 0.05

Figure 8.13 The user selects the strategy to propagate the query within the coalition

The advantage of the two-level propagation is the following: rather than propagating the query throughout the network, the query is sent only to the relevant coalitions.

The propagation graph is displayed as a tree where the databases that were selected are identified (Figure 8.14). The databases that are closer to the graph's root are considered as more relevant, because their context is semantically closer to the query. For example, the DB20 is a database about hydrological disaster in the Quebec province, while DB6 is the Natural Risk Database for Canada, which also contains data on other types of disasters. The propagation graph is a tree because the query propagation algorithm avoid cycles by removing nodes that have already received the query.



The propagation tree shows the partial order of propagation of the query to databases

PROPAGATION ORDER BASED ON QUERY CONTEXT

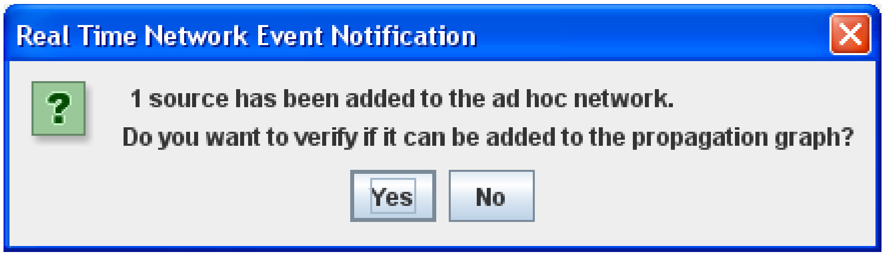
Coalition to Assess Flooding Risk
DB20
DB4
DB23
DB6
DB22

Click Save Propagation Order to save this order for processing query with G-MAP.

Save Propagation Order

Figure 8.14 The user visualizes the propagation graph

Because the ad hoc network is dynamic, new peers can join the network. This creates a problem because if a new source is relevant to answer a query that was sent in a near past, this source will not be considered. Therefore, the user could loose some relevant data. This is what happens in query propagation approaches that are not addressing the real time issue (see Table 6.1). To resolve this problem, the real time query propagation component integrates the real time adaptation algorithm described in Chapter 6, which assesses the need for updating existing propagation graphs when there is a change in the network. First, the user is notified of the arrival of a new peer and is asked if this new source should be considered (Figure 8.15).

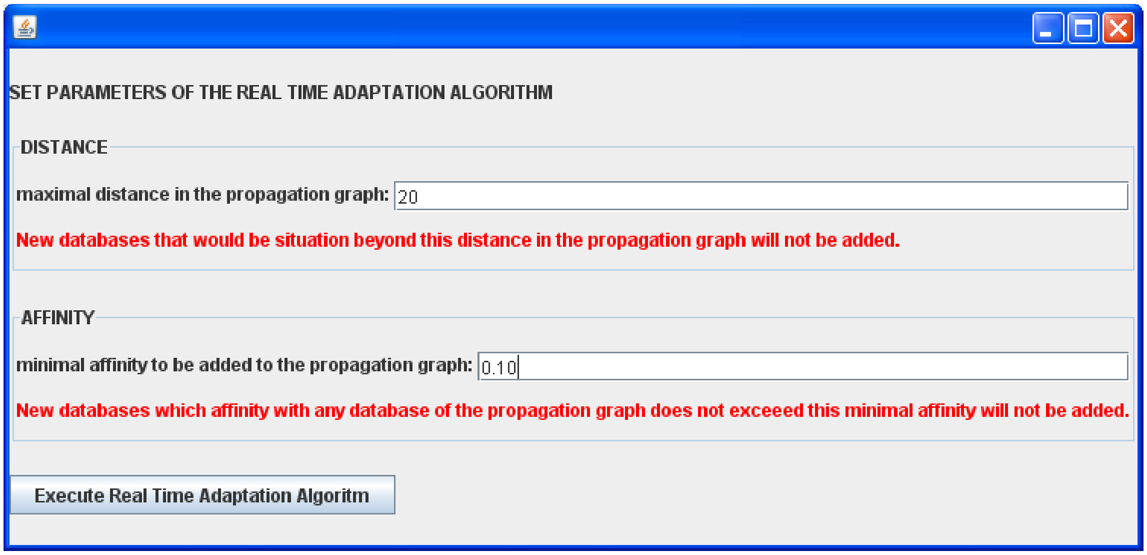


Real Time Network Event Notification

1 source has been added to the ad hoc network.
Do you want to verify if it can be added to the propagation graph?

Figure 8.15 The user is notified of the arrival of a new peer

If the user accepts, he of she is asked to set the parameters of the update algorithm (Figure 8.16).



SET PARAMETERS OF THE REAL TIME ADAPTATION ALGORITHM

DISTANCE

maximal distance in the propagation graph: 20

New databases that would be situation beyond this distance in the propagation graph will not be added.

AFFINITY

minimal affinity to be added to the propagation graph: 0.10

New databases which affinity with any database of the propagation graph does not exceeed this minimal affinity will not be added.

Figure 8.16 The user sets the parameters of the update algorithm

The first parameter is the maximal distance of a peer in the graph. The distance of a peer in the graph is the number of peers that separates it from the root node. The greater the distance, the less relevant the peer is to answer the query. Therefore, if, according to the update algorithm, the new peer is to be placed too far from the root node, it may be irrelevant, because it is semantically too different from the query. The maximal distance sets this distance threshold.

The second parameter is the minimal affinity. Only the new peers whose affinity with a peer of the propagation graph exceeds this threshold will be added to the propagation graph.

The purpose of these two parameters is to determine when a real time update to the propagation is necessary. If the new peer meets these conditions, it is added to the propagation graph. Once the propagation graph is generated, the user in fact obtains a set of peers whose ontologies contain concepts describing data relevant to the query. In the next steps, the aim is to find these concepts. First, the ontologies are semantically augmented with the MVAC component. Then, the semantic reconciliation component finds the concepts that are relevant to the query.

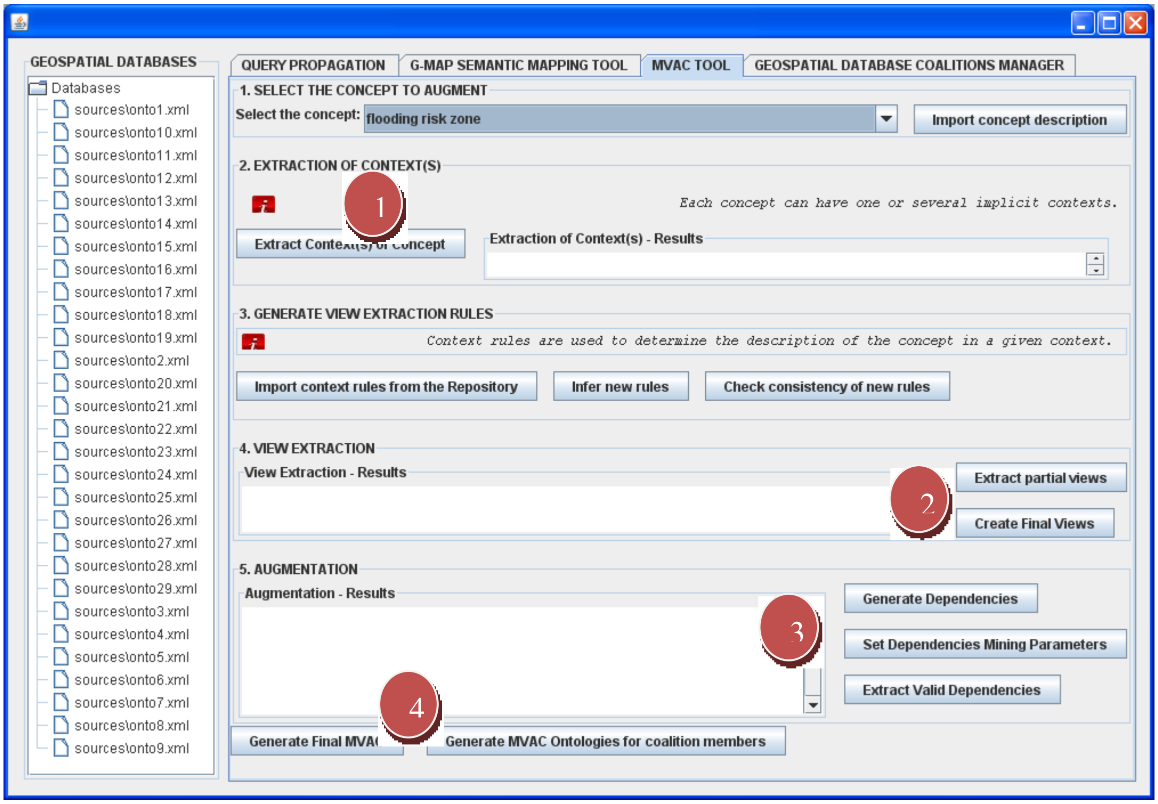
8.6.3 The MVAC Component

When the query concept is compared to concepts of database ontologies, the fact that some semantics may be implicit can refrain from detecting similarities or heterogeneities that are crucial for the user to correctly interpret the data. This component is responsible for enriching the ontologies describing geospatial databases in order to address this issue. The MVAC component can process a set of ontologies and issue augmented ontologies. However, in order to better demonstrate our approach, the component can also process one concept at a time. Figure 8.17 shows the main interface of the MVAC component. The five main services offered by the MVAC component interface are (1) the extraction of contexts of the concept; (2) the view extraction process; (3) the augmentation with dependencies; and (4) the generation of the final MVA concept.

Consider that the user selects the concept “flooding risk zone.” The original description of this concept is the following:

```
FloodingRiskZone ⊑ RiskZone ⊓ (∃ ThP: WaterLevel.low ⊔ ∃ ThP:
WaterLevel.medium ⊔ ∃ ThP: WaterLevel.high) ⊓ (∃ ThP:Status.navigable ⊔ ∃
ThP:Status.NotNavigable) ⊓ (∃ ThP:Risk.low ⊔ ∃ ThP:Risk.medium ⊔ ∃
ThP:Risk.high) ⊓ (∃ STP:SpatialExtent.GML:MovingPolygon) ⊓ (∃
SD:Elevation.low ⊔(∃ SD:Elevation.medium) ⊓ (∃
```

SD:WaterbodyProximity.medium ⊔ ∃ SD:WaterbodyProximity.high) ⊓ (∃ SD:DamProtection.low ⊔ ∃ SD:DamProtection.medium)



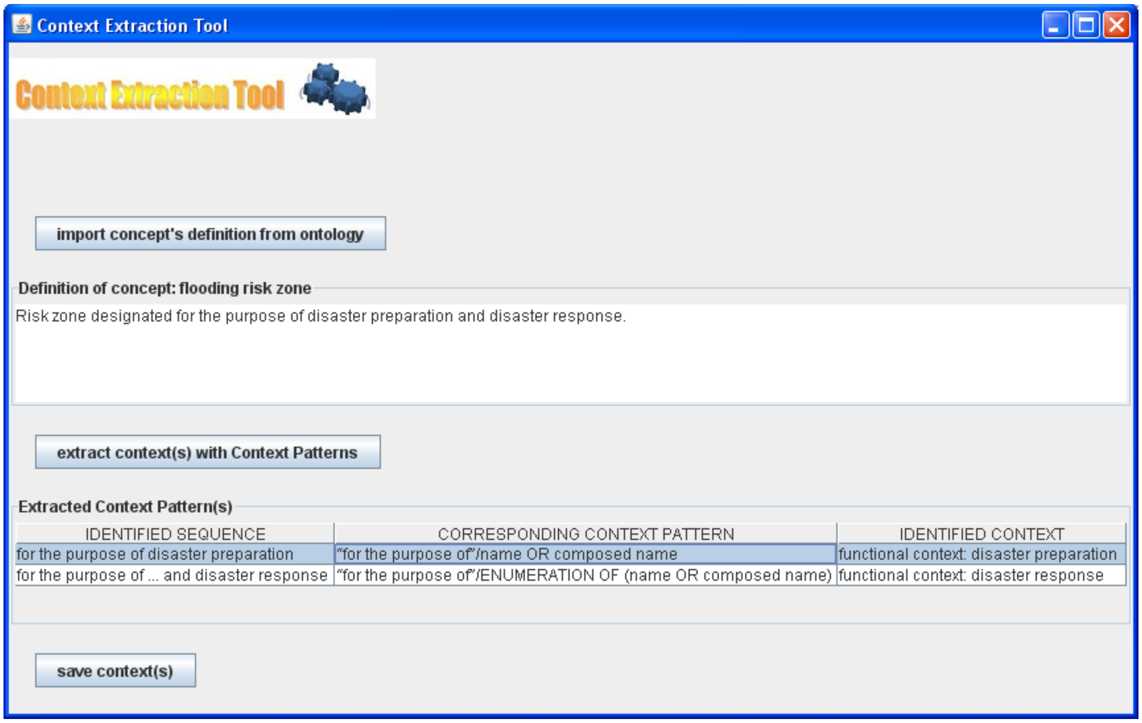
Select the concept: flooding risk zone

Each concept can have one or several implicit contexts.

Context rules are used to determine the description of the concept in a given context.

Figure 8.17 Main interface of the MVAC component

The view extraction engine identifies patterns in the definition of the concept and issues corresponding contexts for the concept (Figure 8.18). For example, the concept “flooding risk zone” has two functional contexts: “disaster response” and “disaster preparation.” It means that flooding risk zones are identified and analyzed either for preparation in case of a disaster, or for reaction during a disaster. In this example, we will see that implicitly, this concept has different meanings depending on the context, and that being aware of these different meanings is very important to ensure that the user can avoid misinterpretation of data.


Context Extraction Tool
Context Extraction Tool
import concept's definition from ontology
extract context(s) with Context Patterns
save context(s)

Definition of concept: flooding risk zone

Risk zone designated for the purpose of disaster preparation and disaster response.

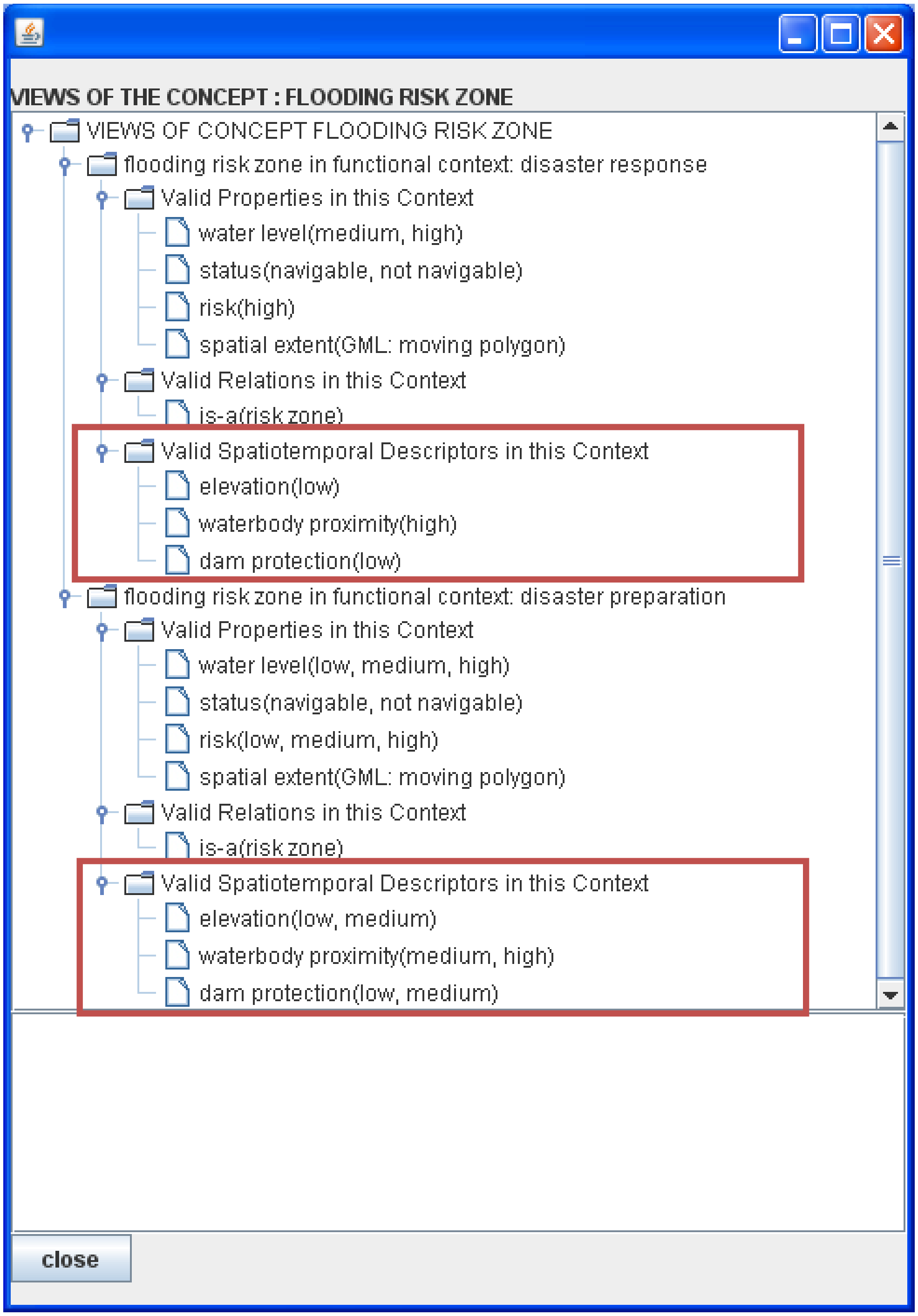
Extracted Context Pattern(s)

IDENTIFIED SEQUENCE	CORRESPONDING CONTEXT PATTERN	IDENTIFIED CONTEXT
for the purpose of disaster preparation	"for the purpose of"/name OR composed name	functional context: disaster preparation
for the purpose of ... and disaster response	"for the purpose of"/ENUMERATION OF (name OR composed name)	functional context: disaster response

Figure 8.18 The user extracts the context of the concept "flooding risk zone"

For each context, a view is extracted from context rules. The user can visualize the values of the concept's features that are valid in each view (Figure 8.19).

In this example, the views differ with respect to the semantics of the spatiotemporal property "*spatial extent*" which is a *GML*: *MovingPolygon*. The semantics of this spatiotemporal property is defined by the spatial descriptors. The spatial descriptors give the factors that characterize a spatial region (the moving polygon) considered as a flooding risk, that is, a low elevation, the spatial proximity to a waterbody and poor dam protection.



VIEWS OF THE CONCEPT : FLOODING RISK ZONE

```
VIEWS OF CONCEPT FLOODING RISK ZONE
  flooding risk zone in functional context: disaster response
    Valid Properties in this Context
      water level(medium, high)
      status(navigable, not navigable)
      risk(high)
      spatial extent(GML: moving polygon)
    Valid Relations in this Context
      is-a(risk zone)
    Valid Spatiotemporal Descriptors in this Context
      elevation(low)
      waterbody proximity(high)
      dam protection(low)
  flooding risk zone in functional context: disaster preparation
    Valid Properties in this Context
      water level(low, medium, high)
      status(navigable, not navigable)
      risk(low, medium, high)
      spatial extent(GML: moving polygon)
    Valid Relations in this Context
      is-a(risk zone)
    Valid Spatiotemporal Descriptors in this Context
      elevation(low, medium)
      waterbody proximity(medium, high)
      dam protection(low, medium)
```

Figure 8.19 The user visualizes the views of the concept “flooding risk zone” in each context. The spatiotemporal descriptors valid in each context are highlighted in boxes.

In the “disaster response” view, the spatial extent of the “flooding risk zone” is defined (in Description Logics syntax) as follows:

```
SpatialExtent⋂∃HasElevation.Low⋂∃WaterbodyProximity.High⋂
∃DamProtection.Low
```

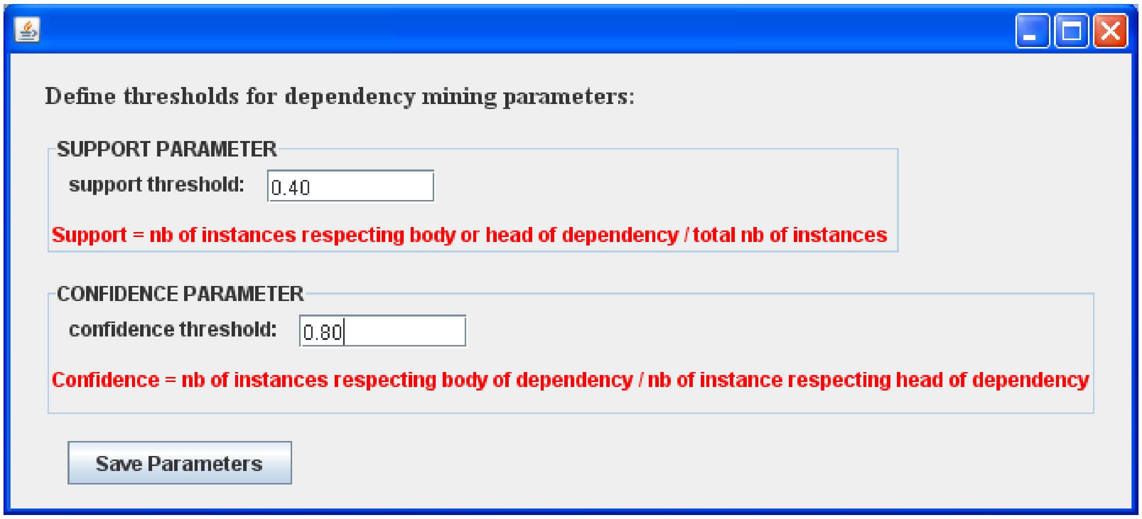
In the “disaster preparation” view, the same spatiotemporal property is defined as:

```
SpatialExtent⋂(∃HasElevation.Low⋃∃HasElevation.Medium)⋂
(∃WaterbodyProximity.High⋃∃WaterbodyProximity.Medium)⋂(∃DamProtection.Low⋃
∃DamProtection.Medium)
```

This means that in the disaster response context, the regions that are considered as “flooding risk zones” must be characterized by a higher absolute level of risk (i.e., the condition to be a risk zone are more restrictive) than in the disaster preparation context. This is because in disaster response, we must focus only on the regions that have the highest risk, whereas in disaster preparation context, we have the time to consider zones with medium risk as well. Therefore, this first semantic enrichment with views allows the user to better interpret the data because the context is no longer implicit. During the semantic mapping process, the user will be allowed, if he or she requests it, to select the context of the query concept that is relevant to his or her situation.

In the next augmentation step, the dependencies between features of the concepts are extracted. To determine the validity of a proposed dependency, it is verified against the instances of the concept, as explained in Chapter 5. To do so, the user is asked to set the thresholds for the confidence and support measures (Figure 8.20).

In this example, the confidence and support were set to 0,80 and 0,40 respectively. Ultimately, the dependencies must be validated by the user, which is expected to have some minimal knowledge of the domain.



Define thresholds for dependency mining parameters:

SUPPORT PARAMETER

support threshold: 0.40

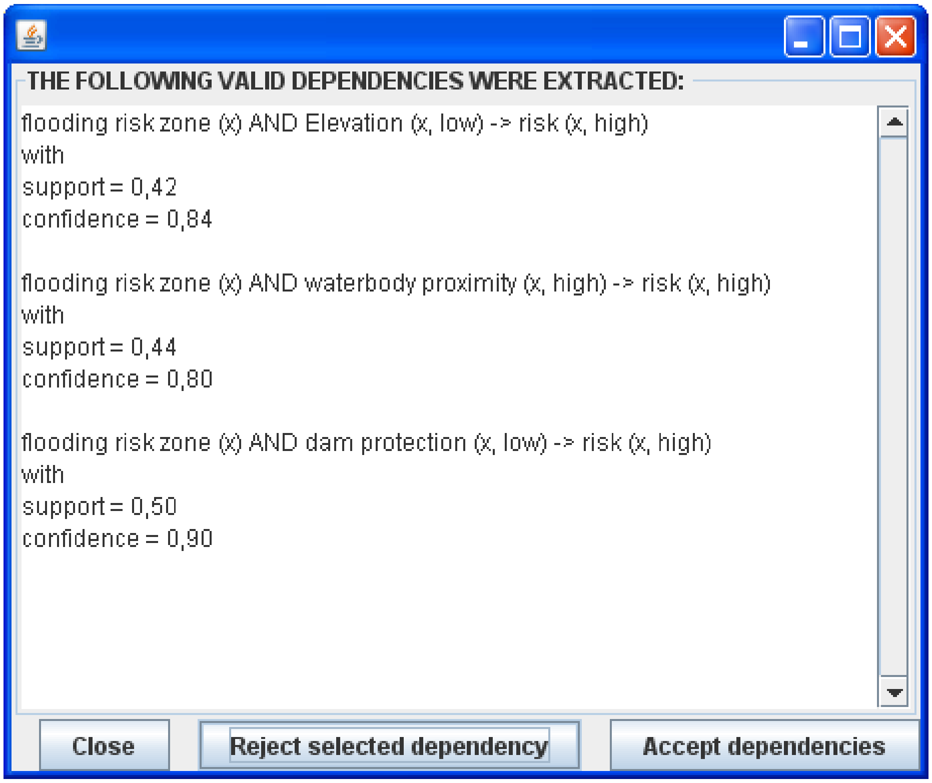
Support = nb of instances respecting body or head of dependency / total nb of instances

CONFIDENCE PARAMETER

confidence threshold: 0.80

Confidence = nb of instances respecting body of dependency / nb of instance respecting head of dependency

Figure 8.20 The user sets the confidence and support thresholds that determine the dependency extraction process



THE FOLLOWING VALID DEPENDENCIES WERE EXTRACTED:

```
flooding risk zone (x) AND Elevation (x, low) -> risk (x, high)
with
support = 0,42
confidence = 0,84

flooding risk zone (x) AND waterbody proximity (x, high) -> risk (x, high)
with
support = 0,44
confidence = 0,80

flooding risk zone (x) AND dam protection (x, low) -> risk (x, high)
with
support = 0,50
confidence = 0,90
```

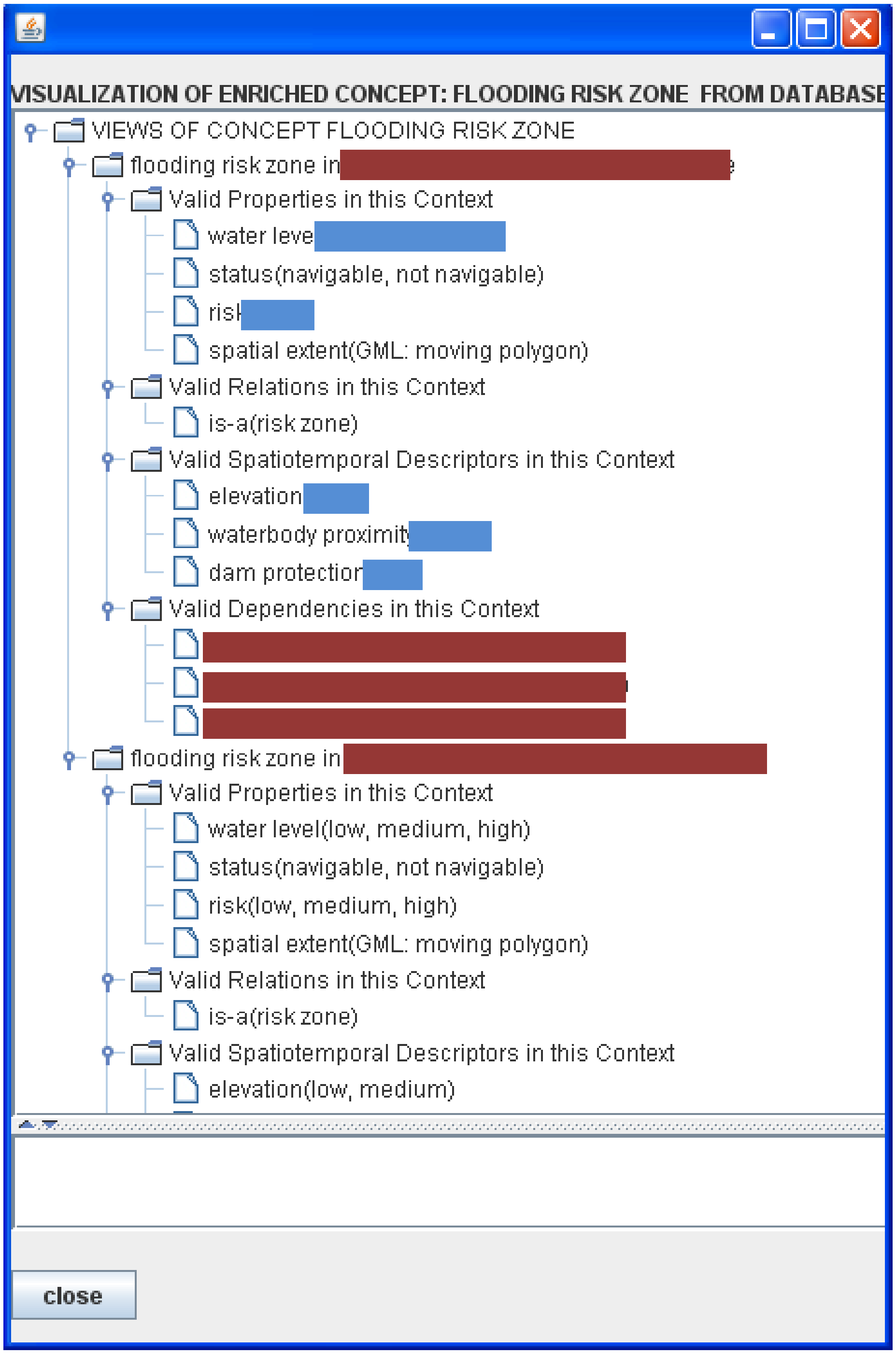
Figure 8.21 The user accepts or rejects the generated dependencies

The following dependencies were identified (Figure 8.21):

flooding risk zone (x) ∧ Elevation (x, low) → risk (x, high)
flooding risk zone (x) ∧ waterbody proximity (x, high) → risk (x, high)
flooding risk zone (x) ∧ dam protection (x, low) → risk (x, high).

The dependencies also allow the user to better understand the links that were implicit between features. In this example, the dependencies could be useful, for example, to assess the flooding risk based on observable properties (elevation, waterbody proximity, dam protection). In the semantic interoperability process, the dependencies will be used to identify “missing” mappings (see next section). Finally, the augmented concept is generated and presented to the user (Figure 8.22). In Figure 8.22, the features that were added are highlighted in red and those that were modified due to enrichment are highlighted in blue to demonstrate the contribution of the MVAC component. The description of the multi-view augmented concept “flooding risk zone” (to be compared with the description of the original concept “flooding risk zone” at the beginning of Section 8.6.3) is the following:

```
mvac: FloodingRiskZone_view1 ⊑ RiskZone ⊓
HasFunctionalContext.DisasterResponse ⊓ (∃ ThP: WaterLevel.medium ⊔ ∃
ThP: WaterLevel.high) ⊓ (∃ ThP:Status.navigable ⊔ ∃
ThP:Status.NotNavigable) ⊓ ∃ ThP:Risk.high ⊓ ∃
STP:SpatialExtent.GML:MovingPolygon ⊓ ∃ SD:Elevation.low ⊓ ∃
SD:WaterbodyProximity.high ⊓ ∃ SD:DamProtection.low,
FloodingRiskZone_view2 ⊑ RiskZone ⊓
HasFunctionalContext.DisasterPreparation ⊓ (∃ ThP: WaterLevel.low ⊔ ∃
ThP: WaterLevel.medium ⊔ ∃ ThP: WaterLevel.high) ⊓ (∃ ThP:Status.navigable
⊔ ∃ ThP:Status.NotNavigable) ⊓ (∃ ThP:Risk.low ⊔ ∃ ThP:Risk.medium ⊔ ∃
ThP:Risk.high) ⊓ ∃ STP:SpatialExtent.GML:MovingPolygon ⊓ (∃
SD:Elevation.low ⊔ ∃ SD:Elevation.medium) ⊓ (∃
SD:WaterbodyProximity.medium ⊔ ∃ SD:WaterbodyProximity.high) ⊓ (∃
SD:DamProtection.low ⊔ ∃ SD:DamProtection.medium),
flooding risk zone (x) ∧ Elevation (x, low) → risk (x, high)
flooding risk zone (x) ∧ waterbody proximity (x, high) → risk (x, high)
flooding risk zone (x) ∧ dam protection (x, low) → risk (x, high).
```



VISUALIZATION OF ENRICHED CONCEPT: FLOODING RISK ZONE FROM DATABASE

- VIEWS OF CONCEPT FLOODING RISK ZONE
  - flooding risk zone in 🟥
    - Valid Properties in this Context
      - water leve 🟦
      - status(navigable, not navigable)
      - risk 🟦
      - spatial extent(GML: moving polygon)
    - Valid Relations in this Context
      - is-a(risk zone)
    - Valid Spatiotemporal Descriptors in this Context
      - elevation 🟦
      - waterbody proximit 🟦
      - dam protectior 🟦
    - Valid Dependencies in this Context
      - 🟥
      - 🟥
      - 🟥
  - flooding risk zone in 🟥
    - Valid Properties in this Context
      - water level(low, medium, high)
      - status(navigable, not navigable)
      - risk(low, medium, high)
      - spatial extent(GML: moving polygon)
    - Valid Relations in this Context
      - is-a(risk zone)
    - Valid Spatiotemporal Descriptors in this Context
      - elevation(low, medium)

Enrichment with new features 🟥

Modification of existing features due to enrichment 🟦

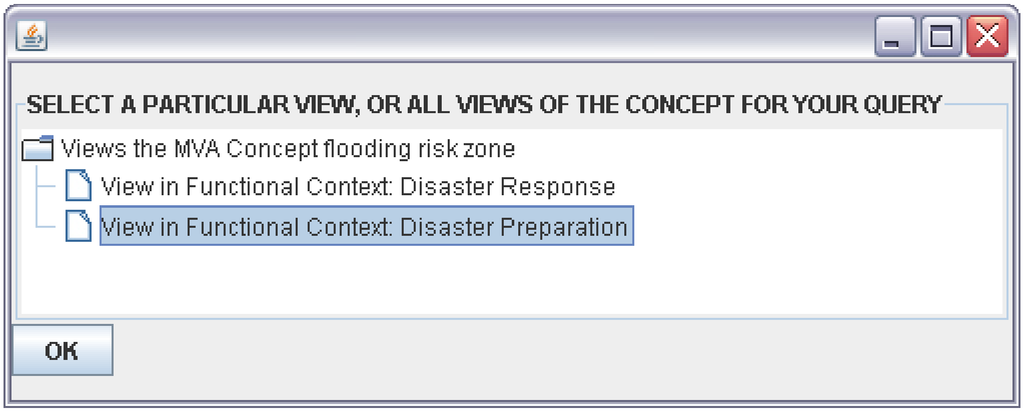
Figure 8.22 The user visualizes the final MVAC

Once the ontologies of selected peers are augmented, the semantic reconciliation component is called to match the query with concepts of the augmented ontologies.

8.6.4 Semantic Reconciliation Component

The semantic reconciliation component integrates the qualitative G-MAP semantic mapping tool and the quantitative Sim-Net semantic similarity tool. The purpose of these tools is to find relevant concepts while identifying semantic heterogeneities, and in particular, the heterogenities that relate to spatial and temporal semantics.

The user first selects the query concept ("flooding risk zone") and the semantic reconciliation component will import the propagation graph that was previously computed for that concept. Then, the user is asked to specify the context that fits his or her situation (Figure 8.23). In the current example, it means that the user chooses the appropriate context based on the definition of flooding risk that best suits his or her needs.



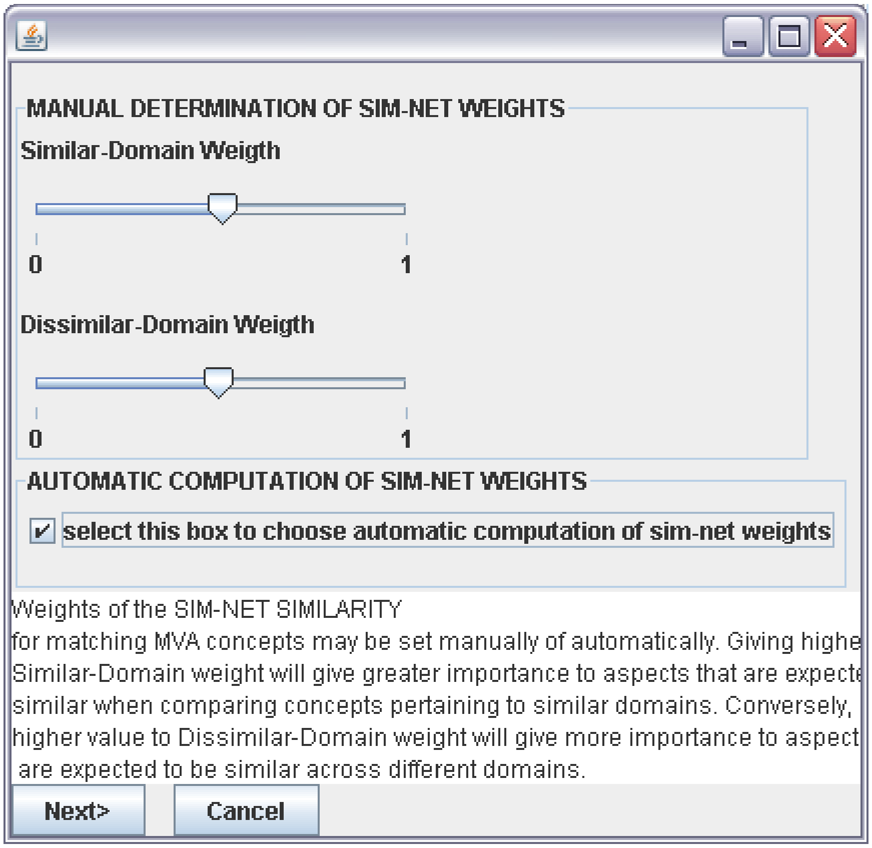
SELECT A PARTICULAR VIEW, OR ALL VIEWS OF THE CONCEPT FOR YOUR QUERY

- Views the MVA Concept flooding risk zone
  - View in Functional Context: Disaster Response
  - View in Functional Context: Disaster Preparation

Figure 8.23 The user specifies the relevant context before computing the semantic similarity

Selecting the context enables to consider only the concept's features that are relevant in that context. For example, if the user selects the context "disaster preparation," a broader range of values will be considered for the properties "water level," "risk," as well as for the spatiotemporal descriptors "elevation," "waterbody proximity," and "dam protection." As a result, the user will be able to retrieve the concepts which have those properties and values during the semantic mapping process.

The user is also asked to specify the weight for the computation of Sim-Net semantic similarity (Figure 8.24). The weights can be set manually or automatically according to the formulas provided in Chapter 7.



MANUAL DETERMINATION OF SIM-NET WEIGHTS

Similar-Domain Weigth

Dissimilar-Domain Weigth

AUTOMATIC COMPUTATION OF SIM-NET WEIGHTS

☑ select this box to choose automatic computation of sim-net weights

Weights of the SIM-NET SIMILARITY
for matching MVA concepts may be set manually of automatically. Giving highe
Similar-Domain weight will give greater importance to aspects that are expecte
similar when comparing concepts pertaining to similar domains. Conversely,
higher value to Dissimilar-Domain weight will give more importance to aspect
are expected to be similar across different domains.

Figure 8.24 The user sets the weights for SIM-NET

The following shows an example of semantic mapping between the query concept “flooding risk zone” in the functional context “disaster response” and the second describing the concept “flood hazard area”, with a single view for simplicity. To identify the types of properties, relations and spatiotemporal descriptors in the MVA concept descriptions, the following tags and their combinations are used: `mvac` for an MVA concept, `Th`, `S` and `T` for thematic, spatial and temporal elements, `P`, `R`, `SD` and `TD` for properties, relations, spatial descriptors and temporal descriptors. The query concept Flooding Risk Zone is defined as:

```
mvac: FloodingRiskZone_view1 ⊑ RiskZone ⊓
HasFunctionalContext.DisasterResponse ⊓ (∃ ThP: WaterLevel.medium ⊔ ∃
ThP: WaterLevel.high) ⊓ (∃ ThP:Status.navigable ⊔ ∃
ThP:Status.NotNavigable) ⊓ ∃ ThP:Risk.high ⊓ ∃
STP:SpatialExtent.GML:MovingPolygon ⊓ ∃ SD:Elevation.low ⊓ ∃
SD:WaterbodyProximity.high ⊓ ∃ SD:DamProtection.low,
```

with the following dependencies:

```
flooding risk zone (x) ∧ Elevation (x, low) → risk (x, high)
flooding risk zone (x) ∧ waterbody proximity (x, high) → risk (x, high)
flooding risk zone (x) ∧ dam protection (x, low) → risk (x, high).
```

The concept Flood Hazard Area is defined as:

```
mvac: FloodHazardArea_view1 ⊑ HazardArea ⊓
HasFunctionalContext.DisasterManagement ⊓ ∃ ThP: WaterLevel.high  ⊓ ∃
ThP:Risk.high ⊓ ∃ STP:SpatialExtent.GML:MovingPolygon ⊓ ∃
SD:GroundLevel.low ⊓ (∃ SD:WaterbodyDistance.low ⊔
SD:WaterbodyDistance.medium),
```

with the following dependency:

```
flood hazard area (x) ∧ ground level (x, low) → risk (x, high)
```

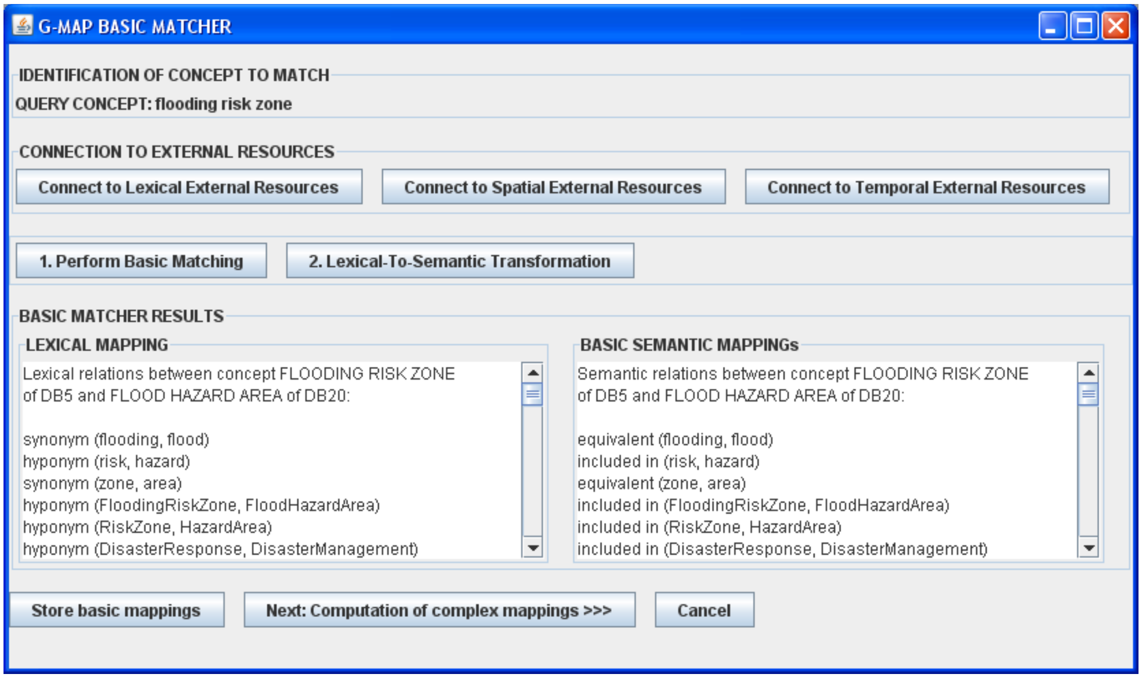
The *Basic MVAC Element Matcher* (Figure 8.25), which uses WordNet and Open Cyc Spatial and Temporal Ontology to identify relations between terms, determines the following non-disjoint basic semantic relations:

```
equivalent (flooding, flood)
included in (risk, hazard)
equivalent (zone, area)
included in (FloodingRiskZone, FloodHazardArea)
included in (RiskZone, HazardArea)
included in (DisasterResponse, DisasterManagement)
```

The G-MAP complex MVAC element mapping inference engine generates (among other) the following composite statement, which verifies the disjoint rule for spatial properties:

```
p(FloodingRiskZone:SpatialExtent) ∧ p(FloodHazardArea:SpatialExtent) ∧
name (FloodingRiskZone:SpatialExtent, SpatialExtent) ∧  name
(FloodHazardArea:SpatialExtent, SpatialExtent) ∧  range
(FloodingRiskZone:SpatialExtent, GML:MovingPolygon) ∧  range
(FloodHazardArea:SpatialExtent, GML:MovingPolygon) ∧  spatial_descriptors
(FloodingRiskZone:SpatialExtent, Elevation) ∧  spatial_descriptors
(FloodHazardArea:SpatialExtent, GroundLevel) ∧  spatial_descriptors
(FloodHazardArea:SpatialExtent, WaterbodyDistance) ∧   disjoint
(Elevation, GroundLevel) ∧ disjoint (Elevation, WaterbodyDistance) ⇒
disjoint (FloodingRiskZone:SpatialExtent, FloodHazardArea:SpatialExtent),
```

meaning that the spatial extent represents different spatial areas that do not overlap, because they have different semantics.


G-MAP BASIC MATCHER
IDENTIFICATION OF CONCEPT TO MATCH
CONNECTION TO EXTERNAL RESOURCES
Connect to Lexical External Resources
Connect to Spatial External Resources
Connect to Temporal External Resources
1. Perform Basic Matching
2. Lexical-To-Semantic Transformation
BASIC MATCHER RESULTS
LEXICAL MAPPING
BASIC SEMANTIC MAPPINGs
Store basic mappings
Next: Computation of complex mappings >>>
Cancel

QUERY CONCEPT: flooding risk zone

Lexical relations between concept FLOODING RISK ZONE of DB5 and FLOOD HAZARD AREA of DB20:

synonym (flooding, flood)
hyponym (risk, hazard)
synonym (zone, area)
hyponym (FloodingRiskZone, FloodHazardArea)
hyponym (RiskZone, HazardArea)
hyponym (DisasterResponse, DisasterManagement)

Semantic relations between concept FLOODING RISK ZONE of DB5 and FLOOD HAZARD AREA of DB20:

equivalent (flooding, flood)
included in (risk, hazard)
equivalent (zone, area)
included in (FloodingRiskZone, FloodHazardArea)
included in (RiskZone, HazardArea)
included in (DisasterResponse, DisasterManagement)

Figure 8.25 The user monitors the computation of basic mappings (lexical relations) and their conversion into semantic relations

Note that at this stage in the reasoning process, the G-MAP was unable to detect that *elevation* and *ground level* represent the same spatial descriptor. In addition, the statements generated by G-MAP, listed below, verify the following non-disjoint mapping rules between thematic properties and relations:

p(FloodingRiskZone:WaterLevel) ∧ p(FloodHazardArea: WaterLevel) ∧ name (FloodingRiskZone:WaterLevel, WaterLevel) ∧ name (FloodHazardArea: WaterLevel, WaterLevel) ∧ range (FloodingRiskZone:WaterLevel, medium ⊔ high) ∧ range (FloodHazardArea: WaterLevel, high) ∧ equivalent (WaterLevel, WaterLevel) ∧ includes (medium ⊔ high, high) ⇒ **includes (FloodingRiskZone: WaterLevel, FloodHazardArea: WaterLevel)**

r(FloodingRiskZone: Is-a) ∧ r(FloodHazardArea: Is-a) ∧ name (FloodingRiskZone: Is-a, Is-a) ∧ name (FloodHazardArea: Is-a, Is-a) ∧ range (FloodingRiskZone: Is-a, RiskZone) ∧ range (FloodHazardArea: Is-a, HazardArea) ∧ equivalent (Is-a, Is-a) ∧ equivalent (RiskZone, HazardArea) ⇒ **equivalent (FloodingRiskZone: Is-a, FloodHazardArea: Is-a)**

p(FloodingRiskZone: Risk) ∧ p(FloodHazardArea: Risk) ∧ name (FloodingRiskZone: Risk, Risk) ∧ name (FloodHazardArea: Risk, Risk) ∧ range (FloodingRiskZone: Risk, high) ∧ range (FloodHazardArea: Risk, high) ∧ equivalent (Risk, Risk) ∧ equivalent (high, high) ⇒ **equivalent (FloodingRiskZone: Risk, FloodHazardArea: Risk)**

The thematic component of the semantic mapping between "*flooding risk zone*" in the functional context "*disaster response*" and the second describing the concept "*flood hazard area*", is given by the following rule of inclusion:

includes (FloodingRiskZone: WaterLevel, FloodHazardArea: WaterLevel) ∧ equivalent (FloodingRiskZone: Is-a, FloodHazardArea: Is-a) ∧ equivalent (FloodingRiskZone: Risk, FloodHazardArea: Risk) ⇒ ∀ f, f∊ FloodingRiskZone_view1_thematic, ∃ f', f'∊ FloodHazardArea_view1_thematic | [equivalent(f, f') ∨ includes (f, f')]⇒ **includes(FloodingRiskZone_view1_thematic, FloodHazardArea_view1_thematic)**

While for the spatial component of the semantic mapping, we obtain:

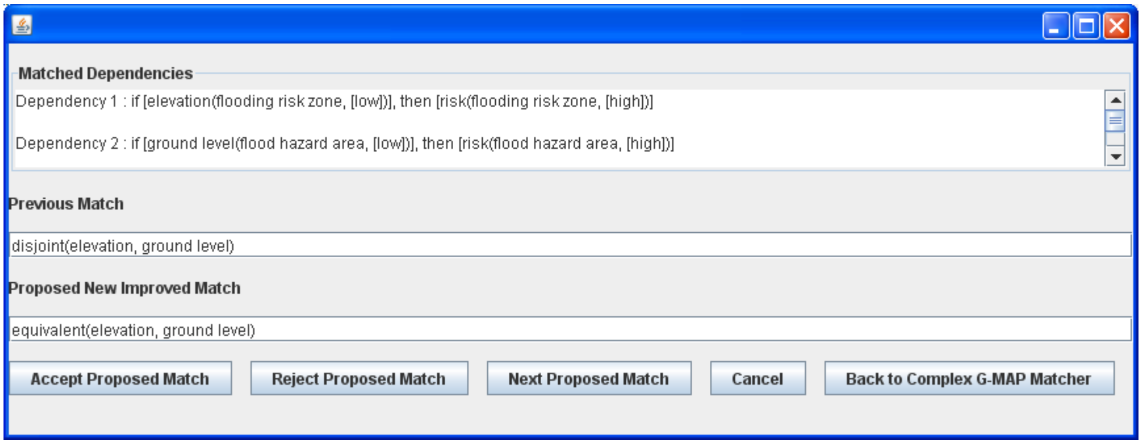
disjoint (FloodingRiskZone: SpatialExtent, FloodHazardArea: SpatialExtent) ⇒ ∀ f, f∊ FloodingRiskZone_view1_spatial, ∀ f', f'∊ FloodHazardArea_view1_ spatial | disjoint (f, f') ⇒ **disjoint (FloodingRiskZone_view1_spatial, FloodHazardArea_view1_ spatial)**

The overall semantic mapping between "*flooding risk zone*" in the functional context "*disaster response*" and the second describing the concept "*flood hazard area*", is inferred by:

```
includes(FloodingRiskZone_view1_thematic, FloodHazardArea_view1_thematic)
∧ disjoint (FloodingRiskZone_view1_spatial, FloodHazardArea_view1_
spatial) ⇒ overlap(FloodingRiskZone_view1, FloodHazardArea_view1)
```

However, the structural correspondence between two dependencies, the first describing the query concept "*flooding risk zone*" and the second describing the concept "*flood hazard area*" of the National Topographic Database (NTDB) respectively, was used by the augmented matcher to infer the correspondence between two properties "*elevation*" and "*ground level*" (Figure 8.26):

```
Equivalent (FloodingRiskZone: Elevation, FloodHazardArea: Ground level)
⇒
(flooding risk zone (x) ∧ Elevation (x, low) → risk (x, high)) ⇔ (flood
hazard area (x) ∧ ground level (x, low) → risk (x, high))
```



Matched Dependencies

Dependency 1 : if [elevation(flooding risk zone, [low])], then [risk(flooding risk zone, [high])]

Dependency 2 : if [ground level(flood hazard area, [low])], then [risk(flood hazard area, [high])]

Previous Match

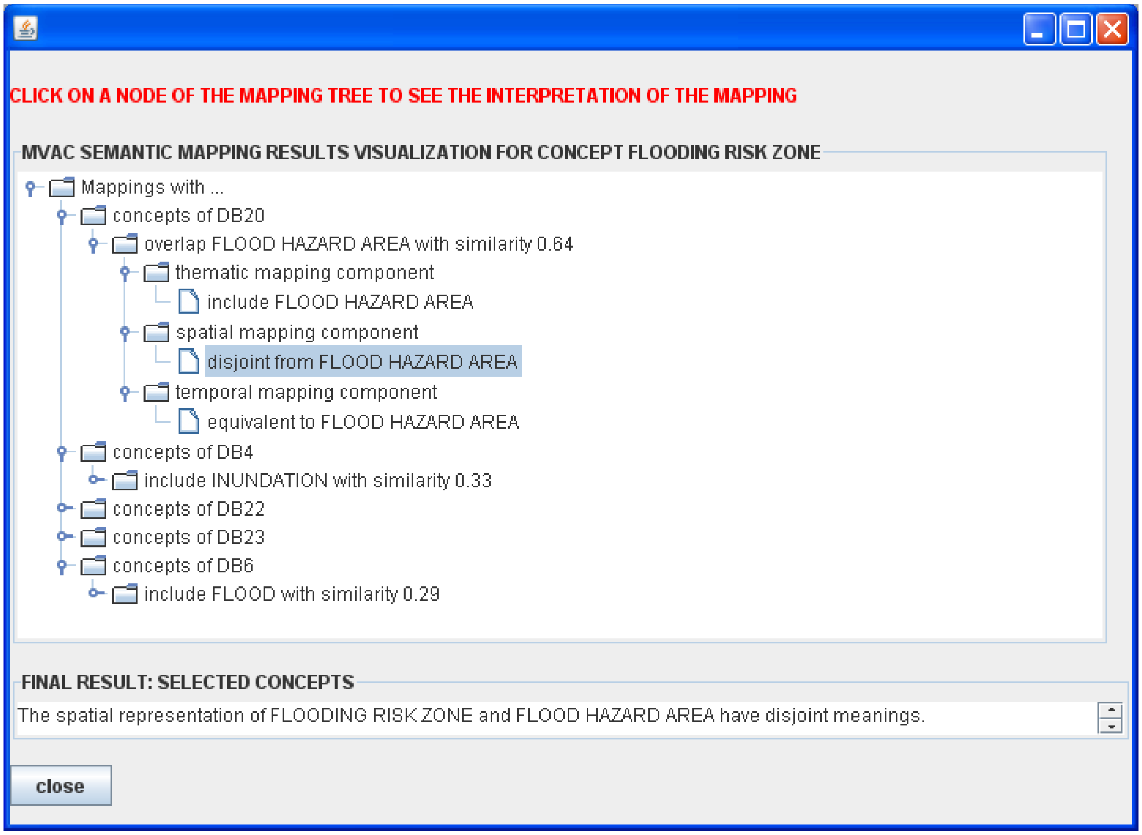
disjoint(elevation, ground level)

Proposed New Improved Match

equivalent(elevation, ground level)

Figure 8.26 Augmented Matcher

Finally, the G-MAP mapping visualization interface shows the concepts that were retrieved by the semantic reconciliation process (Figure 8.27). For each concept that was matched with the query concept "flooding risk zone," the G-MAP mapping system has computed a global semantic relation.



CLICK ON A NODE OF THE MAPPING TREE TO SEE THE INTERPRETATION OF THE MAPPING

MVAC SEMANTIC MAPPING RESULTS VISUALIZATION FOR CONCEPT FLOODING RISK ZONE

```
Mappings with ...
  concepts of DB20
    overlap FLOOD HAZARD AREA with similarity 0.64
      thematic mapping component
        include FLOOD HAZARD AREA
      spatial mapping component
        disjoint from FLOOD HAZARD AREA
      temporal mapping component
        equivalent to FLOOD HAZARD AREA
  concepts of DB4
    include INUNDATION with similarity 0.33
  concepts of DB22
  concepts of DB23
  concepts of DB6
    include FLOOD with similarity 0.29
```

FINAL RESULT: SELECTED CONCEPTS

The spatial representation of FLOODING RISK ZONE and FLOOD HAZARD AREA have disjoint meanings.

Figure 8.27 The user visualizes the answer to the query, i.e. the concepts that match his or her query concept

The novel and distinctive features of the semantic reconciliation component are:

- Hybrid (qualitative and quantitative) semantic mappings that are computed in parallel. With these two complementary pieces of information, the user can make a better decision on which of the proposed concepts is closer to his or her query. For example, even though “flooding risk zone” includes the concept “flood” of DB6, the similarity is 0,29, much lesser than 0.69 similarity with “flood hazard area” of DB20. This means that the concept “flood” might contain much less common features with “flooding risk zone” than “flood hazard area”.

- multi-dimensional semantic mappings, i.e. semantic mapping according to the thematic, spatial and temporal features of the concepts. This allows the user to understand the contribution of each component (spatial, temporal or thematic) in the global semantic relation. For example, the fact that the spatial features of “flooding risk zone” are disjoint from the spatial features of “flood hazard area” might refrain a user who wants to submit a spatial query on flooding risk zones from selecting “flood hazard area”, but not necessarily a user who is only interested by thematic features of the flooding risk zones.

- information on spatiotemporal conflicts: in that case, the definitions of the spatial extent of the risk zone/hazard area are not equivalent. It means that hazard areas in the NTDB are not determined by taking into account the protection of dams, so the NTDB might contain less risk zones. The detection of spatiotemporal conflicts is crucial to avoid misinterpretation of data.

8.7 Discussion and Conclusion

In this chapter, we have presented the prototype that implements our approach for real time semantic interoperability in ad hoc networks of geospatial databases. The prototype implements and integrates four main components: geospatial database coalition discovery, query propagation, MVAC semantic augmentation and semantic reconciliation. The prototype was tested with a set of ontologies that were developed from several database specifications directly or indirectly related to disaster management.

The goal of the implementation and its illustration with the presented case study was to demonstrate that the research hypothesis was valid. Our research hypothesis stated that “it is possible to develop an approach for real time semantic interoperability in ad hoc networks of geospatial databases,” and that “coalition discovery, real time query propagation to relevant sources, semantic augmentation and finally semantic reconciliation enable to achieve real time semantic interoperability in ad hoc networks of geospatial databases.” We also recall from the conceptual framework that we have defined a real-time semantically-interoperable system as “a reactive system that enables the agents to collaborate in order to retrieve and understand shared (geospatial) data they need at run-time." The prototype shows that the combination of these four approaches allows achieving real time semantic interoperability. More specifically, the prototype shows that:

-contexts of databases are useful to discover meaningful coalitions of geospatial databases, and social network analysis techniques are adapted to find relevant coalitions. The coalitions and their contexts are useful for the user who is facing a large set of sources and wants to understand, at a glance, the types of sources contained in the network. Furthermore, coalitions are essential to support effective propagation of queries to relevant sets of sources, without forwarding the query to every source;

- query propagation enables the user to identify the relevant sources in the network, without having to compute semantic mappings between every sources; the query, database and coalition contexts are useful to support this task. Without the query propagation approach, the user would have to access every source of the relevant coalitions to try to identify relevant sources, which is time consuming. In addition, the query propagation approach, in contrast to centralized source discovery approach, can be deployed in a decentralized environment;

- the MVAC semantic augmentation approach allows enriching the ontologies by making explicit the implicit semantics. The MVAC semantic augmentation component makes the user aware of the different contexts of a concept, which supports a better understanding of the data; the formalized representation of spatiotemporal semantics is essential to the understanding of the spatial and temporal extents of real world objects abstracted by concepts, by the user but also by the semantic mapping system;

- the combination of quantitative and qualitative mappings is crucial to support the user in selecting the most relevant concepts with respect to his or her query, since it can be misleading to consider both information separately; dependencies can play a role in the improvement of semantic mapping performance by identifying "missing" mappings, i.e. semantic mappings that could not be identified by lexical or string comparison approaches, nor approaches that use external resources. While dependencies cannot be used as a stand-alone resource to find mappings, they constitute a new way of complementing existing mapping approaches to improve performance. The semantic reconciliation component has the capacity to process complex concept representations, with views, spatiotemporal semantics and dependencies, while existing semantic mapping approaches are restricted to simpler concept representations.

The implementation of the prototype also raised issues that are of interest for future work. For example, it was time consuming to manually produce context rules that support view extraction. It could be investigated how the generation of context rules could be automated with appropriate rule mining techniques. However, since rule-mining techniques use data to discover rules, such approaches would need to be substantially revisited to be adapted to discover rule patterns in database specifications or other relevant source of knowledge. In addition, while the MVAC component aims at enriching concepts by making explicit the implicit semantics (i.e., without external knowledge), it would be important to investigate the role of external resources in such task.

CHAPTER 9

Experimentation

9.1 Introduction

In this section, in addition to the case study which established the feasibility and validity of the approach, we present further experimentation on the prototype. The goal of this experimentation is to show that the proposed approach improves semantic interoperability by enriching the semantics, and exploiting the enriched semantics as much as possible. This experimentation is based on a framework for quality of semantic interoperability, which extends a previous framework on semantic mapping quality published in the *International Symposium for Spatial Data Quality 2007* (ISSDQ 2007) and selected as best paper to be published in *Quality Aspects in Spatial Data Mining* (Bakillah et al. 2009). We have conducted a series of experiments to measure some of the quality characteristics proposed in the semantic interoperability quality framework. The evaluation shows that the different components of the real time semantic interoperability approach for ad hoc networks of geospatial databases contribute to the enrichment and adequate exploitation of semantics, and as a result, semantic interoperability can be improved. Based on the results, we make propositions on how the approach can be further improved.

9.2 Towards a Framework for the Quality of Semantic Interoperability

In this section, we aim to provide the conceptual foundations needed to assess the quality of the semantic interoperability process. The framework for quality of semantic interoperability that we propose is based on a previous framework on the quality of semantic mappings (Bakillah et al. 2009). In previous work (Bakillah et al. 2009), we have introduced the notion of semantic mapping quality and we have conceptualized the elements that affect the process of semantic mapping quality. More specifically, we have indicated that the quality of the semantic mapping process can be assessed based on:

- the quality of the process's input
- the characteristics of the process itself
- the quality of the process's output

Since semantic interoperability can also be described as a process, the conceptualization of our semantic interoperability quality framework is based on an analogous idea. The quality of each semantic interoperability process (coalition formation, semantic augmentation, query propagation or semantic mapping) is described based on the process' characteristics, its input and output. The quality of the global semantic interoperability process is defined based on the quality of its sub-processes.

In this section, we start by explaining the importance of quality assessment for semantic interoperability. Then, we present relevant work that relates to the quality of semantic interoperability. We propose a definition and a model of the quality of semantic interoperability. We develop a set of quantitative measures to evaluate the different characteristics of the quality of semantic interoperability. We demonstrate how this framework can be used to support decision making with a decision matrix. Finally, the framework is employed to evaluate our approach for real time semantic interoperability in ad hoc networks of geospatial databases.

9.3 Why the Quality of Semantic Interoperability is Important?

An important number of semantic interoperability frameworks were developed in recent years, both in the geospatial domain (e.g., Bishr 1998; Brodeur et al. 2003; Kuhn 2003; Lutz et al. 2003; Bian and Hu 2007; Staub et al. 2008; Vaccari et al. 2009; Hossein et al. 2010; Zhang et al. 2010) and in the larger information system community (e.g., Park and Ram 2004; Cudré-Mauroux 2006; Keeney et al. 2006). Those frameworks aim to support the integration of multiple databases, to enable communication between different software agents, to integrate environmental process models, or to enable the discovery of relevant sources or geospatial services. The quality of the semantic interoperability process affects the decision-making process that pertains to all of these tasks. Various aspects may influence and lower the quality of semantic interoperability. For example, poor semantics impedes the user from understanding the meaning of shared data; a semantic mapping process that does not take into account all the features of compared concepts will not detect some semantic heterogeneities. As another example, semantic mappings are used to translate a query submitted to the global schema of a federated database system into a query on the local schema of a local database (approach Global-as-View, GAV) or, conversely, to translate the query submitted to the schema of a local database into a query on the global schema (approach Local-as-View, LAV). Therefore, the quality of semantic mappings impacts the quality of answers to queries processed over multiple sources. When reusing data, users make decisions based on the meaning of retrieved data. Therefore, if the semantics of shared data is affected by the semantic interoperability process, the decision-making process will also be affected. Conversly, if users are aware of the quality of the semantic interoperability process, they can make sound decisions. A semantic interoperability quality framework can assist in the interpretation of the results and reduce false interpretation.

Assessing the quality of the semantic interoperability process is therefore an important issue that must be addressed in parallel with the development of semantic interoperability frameworks.

The notion of data quality has been extensively studied, whether it refers to internal quality (meaning the absence of errors in data) or external quality (how data fit the user's needs, or “fitness-for-use”) (Couclelis 2003; Devillers et al. 2007; Goodchild 2007; Curé and Jeansoulin 2007; Congalton and Green 2009; Sadiq and Duckam 2009). Several quality modeling and assessment approaches assume that data is accessed centrally; in this context, the quality of query results corresponds to the quality of the retrieved data (Zaihrayeu 2006). However, to the best of our knowledge, none of the existing work define the quality of semantic interoperability in decentralized settings such as ad hoc networks, and no framework on the quality semantic interoperability have been proposed yet. Nevertheless, some related research has been proposed; it is presented in the following section.

9.4 Related Work

According to Zaihrayeu (2006), in a distributed environment such as peer-to-peer (P2P) networks, users cannot expect correct and complete query results, but they have to accept incomplete and partially incorrect results. He proposes the notion of *good-enough answers* to assess the quality of query answers in P2P information systems, and more particularly during query propagation. A good-enough answer is defined as "*an answer to a user query which serves its purpose given the amount of effort made in computing it*" (Giunchiglia and Zaihrayeu 2002). Aberer et al. (2003) have studied the impact of the quality of semantic mapping in a P2P query propagation setting. They focus on the specific case where a query being propagated in a P2P network reaches the same peer after going through a loop. They argue that the query, which is translated at each "hop" using semantic mappings between two nodes, should be expressed with the same initial concepts and attributes taken from the schema of the peer after going through the loop (as in the "grapevine telegraph" or "broken telephone"). If this condition is not verified, it means that some of the semantic mapping(s) in the loop are of low quality. In order to assess the quality of a query answer, Yang and Garcia-Molina (2002) proposed to use the number of query answers, and the measure of time between the moment of submission of the query and the moment when a minimal number of query answers is received. Such indicators do not take into account the semantic aspects of quality. Löser et al. (2003) have developed a set of quality dimensions, including completeness, accuracy, time to respond and amount of data, in order to assess the quality of semantic overlay (the semantic layer over the network) of P2P networks. In 2009, Mochol indirectly addressed the problem of semantic mapping quality; she has developed a tool that helps the user to select the most appropriate matching algorithm considering a given matching task. In the Metadata-based Ontology Matching (MOMA) Framework, she has modeled with dependency rules the suitability of existing matching algorithms to match certain types of input (ontologies). In order to determine which matchers are appropriate for a given matching task, the MOMA Framework compares metadata on ontology with metadata on available matchers. The expected result is that quality of matching is optimal when the most suitable matching tool is employed. Other methods to assess the quality of semantic mappings do exist, but they focus on a global performance evaluation of the semantic mapping process, generally by using precision and recall metrics, as well as the f-measure and overall-measure (Do et al. 2003). These metrics are based on a set of reference mappings (the "real" correspondences), which are manually identified by experts. Therefore, they cannot indicate the quality of semantic mappings when no reference is available.

A common emerging agreement that appears to be shared by the above approaches is that a variable quality of results is intrinsic to semantic interoperability processes (query propagation, semantic mapping, etc.). Therefore, when quality cannot be improved, it must at least be monitored and communicated to the user in a meaningful manner to support him or her in making a sound decision (regarding which tool to use, which query results to select, etc.). Nevertheless, establishing the foundation of what is the quality of semantic interoperability is still a strong requirement that was not fulfilled as of today. In order to assess the quality of semantic interoperability, we propose a framework that includes a set of quality indicators and measurements for these indicators. Prior to presenting those indicators, we propose a definition of semantic interoperability quality.

9.5 How to Define Quality of Semantic Interoperability in Ad Hoc Networks of Geospatial Databases?

A discussion regarding the quality of semantic interoperability in ad hoc networks of geospatial databases must begin with a definition of this concept. According to the ISO 9000 standard, quality is "*the totality of the properties and characteristic of a product or service which influence its ability to satisfy explicit or implicit needs*" (ISO Standards 9000, 2000). Following this statement, we propose the following definition:

Definition 9.1: Quality of Semantic Interoperability

Semantic interoperability quality is the totality of the properties and characteristics of the components, processes, input and output that influence the ability of the interoperable system to deliver reusable, relevant and meaningful information to users of this interoperable system and to satisfy their explicit needs. The term *reusable* means that the data that is delivered to a user by the interoperable system can be used to perform the tasks the user intend to perform, which may be different from the tasks that the data was originally intended for. The term *relevant* means that the data delivered to the user corresponds to his or her information needs. The term *meaningful* means that the user is able to interpret the data, that is, to understand what the data refers to in reality. This definition suggests that a model for quality of semantic interoperability will represent the quality of all sub-processes involved in the global semantic interoperability process.

9.6 Model for the Quality of Semantic Interoperability

The Model for the Quality of Semantic Interoperability that we have developed is based on the principle that we can determine the quality of each individual process that participates in the global semantic interoperability process, in order to gradually monitor its quality. The model currently proposed in this thesis represents the semantic interoperability processes that compose the proposed framework for real time semantic interoperability in ad hoc networks. However, this framework for quality of semantic interoperability is flexible enough to include other processes that could play a role in the global semantic interoperability process.

The proposed model extends the model for Quality of Semantic Mappings that was proposed in (Bakillah et al. 2009). It is based on the same fundamental idea that the quality of a process depends on the quality of the input (QofI), the quality of the process itself (QofP) and the quality of the output (QofO). The quality of the global process is gradually monitored through the quality of the individual processes. Therefore, we define the gradual quality of semantic interoperability as “a quality that can be monitored by the user as the global semantic interoperability process unfolds throughout the various phases, so that the user can make appropriate decisions at each phase of the global process.” Figure 9.1 illustrates this principle.

The model for quality of semantic interoperability is illustrated with an UML schema in Figure 9.2. The objective of this conceptual model is not to represent the process for assessing the quality of semantic interoperability, but rather to represent the classification of the different quality characteristics.

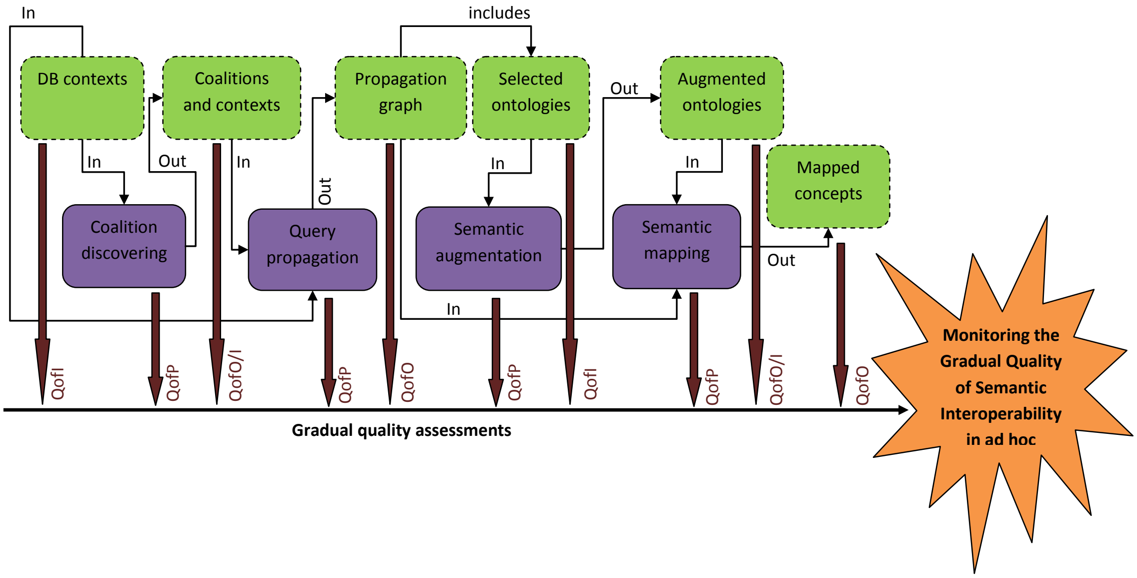


Figure 9.1 Principle for assessing the quality of semantic interoperability

The gradual semantic interoperability quality is a function of (1) the quality of coalition discovery; (2) the quality of semantic augmentation; (3) the quality of query propagation; and (4) the quality of semantic mapping between concepts of ontologies. The methodology that we have adopted is the following: first, we establish that the quality of any semantic interoperability process (hereafter denoted as SIP) depends on the quality (1) of its input; (2) of the process itself, and (3) of its output. We provide eight general quality characteristics, which are classified under one of these three categories. Finally, we explain how these quality characteristics are interpreted for each of the four SIPs. We note that the set of quality characteristics is not intended to include all possible quality characteristics, but intends to provide a basis for the development of a fully comprehensive framework.

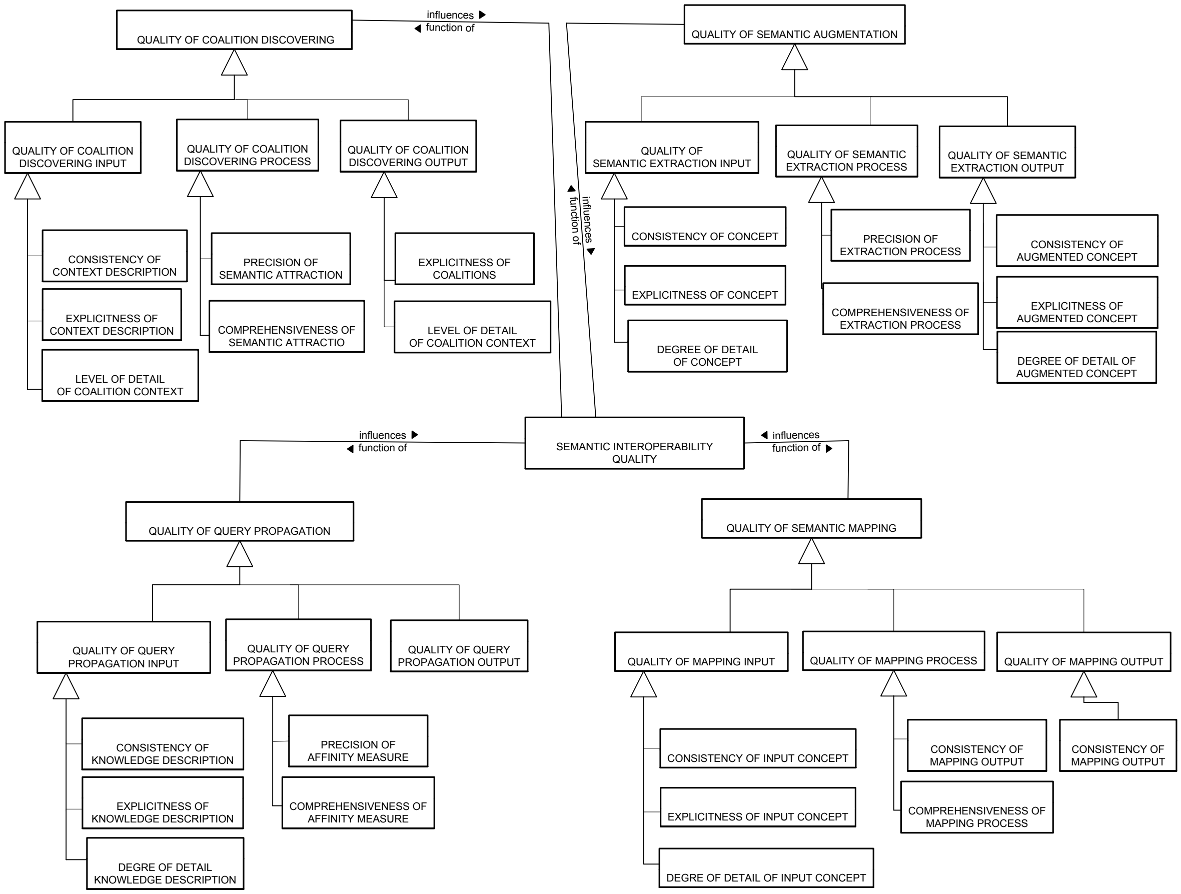

QUALITY OF COALITION DISCOVERING
influences ▶
◀ function of
QUALITY OF SEMANTIC AUGMENTATION
QUALITY OF COALITION DISCOVERING INPUT
QUALITY OF COALITION DISCOVERING PROCESS
QUALITY OF COALITION DISCOVERING OUTPUT
QUALITY OF SEMANTIC EXTRACTION INPUT
QUALITY OF SEMANTIC EXTRACTION PROCESS
QUALITY OF SEMANTIC EXTRACTION OUTPUT
◀ function of
influences ▶
CONSISTENCY OF CONTEXT DESCRIPTION
EXPLICITNESS OF CONTEXT DESCRIPTION
LEVEL OF DETAIL OF COALITION CONTEXT
PRECISION OF SEMANTIC ATTRACTION
COMPREHENSIVENESS OF SEMANTIC ATTRACTIO
EXPLICITNESS OF COALITIONS
LEVEL OF DETAIL OF COALITION CONTEXT
CONSISTENCY OF CONCEPT
EXPLICITNESS OF CONCEPT
DEGREE OF DETAIL OF CONCEPT
PRECISION OF EXTRACTION PROCESS
COMPREHENSIVENESS OF EXTRACTION PROCESS
CONSISTENCY OF AUGMENTED CONCEPT
EXPLICITNESS OF AUGMENTED CONCEPT
DEGREE OF DETAIL OF AUGMENTED CONCEPT
influences ▶
◀ function of
SEMANTIC INTEROPERABILITY QUALITY
◀ influences
function of ▶
QUALITY OF QUERY PROPAGATION
QUALITY OF SEMANTIC MAPPING
QUALITY OF QUERY PROPAGATION INPUT
QUALITY OF QUERY PROPAGATION PROCESS
QUALITY OF QUERY PROPAGATION OUTPUT
QUALITY OF MAPPING INPUT
QUALITY OF MAPPING PROCESS
QUALITY OF MAPPING OUTPUT
CONSISTENCY OF KNOWLEDGE DESCRIPTION
EXPLICITNESS OF KNOWLEDGE DESCRIPTION
DEGRE OF DETAIL KNOWLEDGE DESCRIPTION
PRECISION OF AFFINITY MEASURE
COMPREHENSIVENESS OF AFFINITY MEASURE
CONSISTENCY OF INPUT CONCEPT
EXPLICITNESS OF INPUT CONCEPT
DEGREE OF DETAIL OF INPUT CONCEPT
CONSISTENCY OF MAPPING OUTPUT
COMPREHENSIVENESS OF MAPPING PROCESS
CONSISTENCY OF MAPPING OUTPUT

Figure 9.2 Model for the Quality of Semantic Interoperability

The quality of input refers to the quality of the knowledge representations (KR) that are exploited by the SIPs. The possible KRs include database contexts, ontologies, rules, concept definitions, etc. The quality-of-input characteristics include consistency, explicitness, and level of detail.

The quality of the process refers to the adequacy of the process with respect to the input KR. If the process is not adapted to the input, it means that all the available information is not exploited. The quality-of-process characteristics include precision and comprehensiveness.

Finally, the quality of the output refers to the quality of the new KRs produced by the SIPs. These KRs include coalitions of databases, augmented ontologies, propagation graphs, and semantic mappings. The quality-of-output characteristics include the consistency, the explicitness and the level of detail of the output.

In the following, we define the eight quality characteristics that we have developed. In Section 9.7, more details will be given on how to concretely interpret and assess these characteristics.

Definition 9.2: Consistency of the input. This characteristic assesses the agreement between the current KRs and the constraints that are imposed on this type of KR. For example, within the database's context, a constraint could be that the geographical location should correspond to a location identified in a designated ontology of places. Such constraints are necessary to ensure that the SIP has the capacity to process the input KRs.

Definition 9.3: Explicitness of the input. This characteristic assesses the richness of the KRs. More specifically, it assesses the presence of expected features in the input KRs. For example, it is expected in the MVA concept's definition that each spatial property is associated with a spatial descriptor. If some expected features are missing in an input KR, the explicitness of the input is reduced.

Definition 9.4: Level of detail of the input. This characteristic assesses the level of detail of the input KR elements. The definition of the level of detail depends on the nature of the KR element. For example, in the databases' contexts, the temporal period can be expressed only with the year, or contain more level of detail such as the month, day, hour. etc. Similarly, the role of the database can be very broad, for example "flood monitoring," or it can be more specific, for example "measure water level." A low level of detail will result in limited information for the user to interpret shared data; conversely, a high level of detail ensures that the user can correctly understand the nature of the data.

Definition 9.5: Precision of the process. This characteristic assesses the adequacy of the SIP with respect to the level of detail of the input. For example, if the process has the capacity to compare only temporal periods expressed as years, but not at a greater level of granularity, the precision of the process with respect to the precision of the input is low. The precision of the SIP is therefore a relative measure. A low precision process affects the global SIP because it means that the semantic richness of KRs is not exploited. Therefore, a low precision process results in semantic loss.

Definition 9.6: Comprehensiveness of the process. This characteristic assesses the adequacy of the SIP with respect to the explicitness of the output. It measures the ability of the SIP to take into account all of the types of features of the input KRs. For example, a semantic mapping process that does not have the capacity to compare spatial and temporal descriptors (which are features of the augmented concept) has a low comprehensiveness. A process with low comprehensiveness also results in semantic loss.

Definition 9.7: Consistency of the output. This characteristic assesses the level of conflict between the semantic relations (semantic mappings) that are computed during a SIP. For example, let's suppose that we have three concepts $\{a_0, a_1, a_2\}$ and $\{b_0, b_1, b_2\}$ respectively from ontologies *A* and *B*. Let's consider that within their respective ontologies, these concepts are linked with the following subsumption relationships: $a_0 > a_1 > a_2$, and $b_0 > b_1 > b_2$, where $a_o > a_1$ indicates that a_o subsumes a_1 (or a_0 is a generalization of a_1). Moreover, let's consider that the following semantic mappings were computed: $a_1 < b_1$ and $a_1 > b_0$. These mappings are conflicting, since we can deduce from them that $b_0 < b_1$, which contradicts the relationships stated in ontology *B*. In (Bakillah et al. 2009), we have established a set of conditions, called Mapping Conflict Predicates (MCP), which express these kinds of conflicts between mappings. This measure can be applied only to the SIPs that involve a semantic mapping process which issues set-based qualitative relations.

Definition 9.8: Explicitness of the output. This characteristic is analogous to the explicitness of the input. It assesses the presence or absence of expected features in the output KRs. For example, it is expected that the MVAC augmentation component produces augmented concepts with views for the different extracted contexts. If no views could be extracted, then the level of explicitness was not improved. Consequently, if some of the expected features are missing in an output KR, this will have a negative impact on the the explicitness of the output.

Definition 9.9: Level of detail of the output. This characteristic is analogous to the level of detail of the input. It assesses the level of detail of the output KR elements; the level of detail amounts to the information content of the input element. For example, if we consider the level of detail of a coalition's context, it depends on the level of detail of the role, of the domain, of the geographical location and of the temporal period.

The interpretation of those quality characteristics for the four SIPs considered in this thesis is provided in Table 9.1.

Table 9.1 Interpretation of the Quality Characteristics for the Different Semantic Interoperability Processes

Quality Characteristics		Coalition Discovery process	Semantic Augmentation process (MVAC)	Semantic Mapping Process	Query Propagation process
Quality of Input	Consistency of input	A measure of the agreement between constraints on databases' contexts and the current input contexts	A measure of the agreement between constraints in the ontology and the current input ontology specification	A measure of the agreement between constraints in the ontology and the current input ontology specifications	A measure of the agreement between the constraints on input KRs (databases' and coalitions' context, queries) and the current input KRs
	Explicitness of input	A measure of the number of expected database's context features that are present in the current input contexts	A measure of the number of expected ontological features that are present in the current input ontologies	A measure of the number of expected ontologies' features that are present in the current input ontologies	A measure of the number of expected KR features that are present in the current input KRs
	Degree of detail of input	A measure of the degree of detail of the input context features	A measure of the degree of detail of the input ontology features	A measure of the levele of detail of the input ontologies' features	A measure of the level of detail of the input KR' features
Quality of Process	Precision of process	Agreement between the precision of the semantic attraction measure and the precision of input databases' contexts features	Agreement between the precision of the augmentation processes (context extraction, view extraction, dependency extraction) and the level of detail of ontology's features	Agreement between the precision of the semantic mapping processes and the level of detail of input ontology's features	Agreement between the precision of the semantic affinity measure and the input KR precision

	Comprehensiveness of process	Agreement between the set of features that are compared by the semantic attraction measure and the set of features of input databases' contexts	Agreement between the set of features that are taken by the augmentation processes (context extraction, view extraction, dependency extraction) and the set of features of input ontology	Agreement between the set of features that are taken by the semantic mapping process and the set of features of input ontologies	Agreement between the set of features that are compared by the semantic affinity measure and the set of features of input KRs
Quality of Output	Consistency of output	does not apply	does not apply	A measure of the number of MCPs that are verified by the semantic mappings between concepts	does not apply
	Explicitness of output	A measure of the number of expected coalition's context features that are present in the current output coalitions' contexts	A measure of the number of expected ontological features that are present in the augmented ontologies	A measure of the number of expected mapping features that are present in the output mappings	A measure of the number of expected features that are present in the output propagation graph
	Degree of detail of output	A measure of the degree of detail of the output coalitions' context features	A measure of the degree of detail of the augmented ontology features	does not apply	does not apply

In the following, we propose quantitative measures to assess the proposed quality characteristics.

9.7 Measures to Assess the Quality of Semantic Interoperability

In this section, we provide the mathematical functions that measure the quality characteristics. Each function returns a value between 0 and 1, where 0 indicates lowest quality and 1 indicates highest quality. The measures are defined depending on the user's expectations in terms of quality. Each formula is generic, i.e., not specific to a SIP in particular. Hence, the formulas are adaptable to other types of SIP that were not considered in this thesis but that may participate in the global semantic interoperability process. It is explained how each formula can be used to assess the quality of the different SIP by a simple adaptation of the generic formula to the relevant variables.

9.7.1 Measuring Consistency of Input

The *consistency of input* measures the ratio of the number of input features that respects some constraints (regarding format or cardinality, for example) to the number of features that do not respect these constraints. Consider a constraint *i*, and let $El^t_{\text{total}}(i)$ be the total number of input features of type *t* that should respect the constraint *i*. For example, *t* could be the role, the domain, the geographical location, or the temporal period of a database's context. Let $El^t_{\text{verify}}(i)$ be the total number of input features of type *t* that verify this constraint. The consistency of input, with respect to this constraint, is definedby the following ratio:

$$\text{consistency of input for constraint i} = \frac{El^t_{verify}(i)}{El^t_{total}(i)}$$

Consider that there are several constraints, say n_t, that affect a type of input feature *t* (ex: all constraints affecting the geographical location in a context description). To obtain the global consistency for a type *t* of input feature, we sum all consistency ratios for each constraint *i* and normalize with the number of constraint n_t:

$$\text{consistency of input for a single type t of input feature} = \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{El^t_{verify}(i)}{El^t_{total}(i)}$$

Finally, consider that the global input of a SIP is composed of *m* types of input features (ex: context is composed of a role, domain, etc.). To obtain the global consistency of the SIP input, we sum the values of consistency for all types of input features and normalize with the number of types of input features *m*. As a result, the formula for the consistency of the input of a SIP is

$$\text{consistency of input} = \frac{1}{m}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{El^1_{verify}(i)}{El^1_{total}(i)} + \frac{1}{n_2}\sum_{i=1}^{n_2}\frac{El^2_{verify}(i)}{El^2_{total}(i)} + \dots\right]$$

where t = 1, …*m*. The consistency of input is a value situated between 0 (for low consistency) and 1 (for high consistency). The formula can be extended for an arbitrary number *m* of types of input features.

9.7.2 Measuring the Explicitness of Input

The *explicitness of input* measures the ratio of the number of input features that are taken by a SIP to the number of expected features. For example, the expected features of a database's context include at least one domain, one function, a geographical location and a time period. Let $\mathrm{El}^{i}_{\mathrm{expect}}$ be the number of expected features (in the previous example, $\mathrm{El}^{1}_{\mathrm{expect}} = 4$), and $\mathrm{El}^{1}_{\mathrm{present}}$ the number of present features in a single input element *i* (for example, a single context *i*). The explicitness of this input element *i* is defined by the following ratio:

$$\text{explicitness of input for a single input element} = \frac{El^{i}_{present}}{El^{i}_{\mathrm{expect}}}$$

If we consider an arbitrary number *m* of input elements (for example, the coalition discovering process takes as input *m* databases' contexts), the global explicitness of input is given by their sum, normalized with *m*:

$$\text{explicitness of input} = \frac{1}{m}\left[\frac{El^{1}_{present}}{El^{1}_{\mathrm{expect}}} + \frac{El^{2}_{present}}{El^{2}_{\mathrm{expect}}} + \ldots + \frac{El^{m}_{present}}{El^{m}_{\mathrm{expect}}}\right]$$

The number of expected features is set by the conceptual model for the input KR, but it can be fine-tuned according to the user's requirements. The explicitness is fundamental because low explicitness means that the comparison of knowledge representations is based on incomplete information. It is expected that semantic augmentation can improve explicitness of concepts. However, the formula for explicitness depends on the expected number of elements. This makes this measure flexible for different situations where different levels of explicitness are required.

9.7.3 Measuring Level of Detail of Input

The *level of detail of input* measures the level of detail of a SIP input feature, with respect to the expected level of detail for this feature. The measure for the level of detail is similar to the measure for explicitness, since it also compares an expected level of detail for an input feature *i*, denoted $\mathrm{LoD}^{i}_{\mathrm{expected}}$, with the current level of detail, denoted $\mathrm{LoD}^{i}_{\mathrm{current}}$. The level of detail of each input is added for each of the *m* SIP input features:

$$\text{level of detail of input} = \frac{1}{m}\left[\frac{LoD^{1}_{current}}{LoD^{1}_{\mathrm{exp}ect}} + \frac{LoD^{2}_{current}}{LoD^{2}_{\mathrm{exp}ect}} + \ldots + \frac{LoD^{m}_{current}}{LoD^{m}_{\mathrm{exp}ect}}\right]$$

The LoD measure is flexible, since the $\mathrm{LoD}_{\mathrm{expected}}$ can be set according to the user's requirements. But as for the expected level of explicitness, it can be naturally set according to the conceptual model of the input KR. For example, it can correspond to the expected granularity of the time period in a database's context description (year, month, day, hour. etc.), where the level of detail is quantified according to the hierarchy of time units. For a spatial input feature such as the geographical localisation specified in a database's context, it can correspond to the expected granularity of spatial subdivision (continent, country, province/state, city, community). For a thematic input feature, such as the name of a concept or the domain specified in a database's context, it can correspond to the level where the feature is situated in the hierarchy defined in the global ontology being used to support the identification of related concepts. These cases are illustrated in Figure 9.3.

The consistency, explicitness and level of detail of input are measures which aim to inform the user about the quality of the knowledge representations used by the SIPs. The following quality characteristics measure the quality of the processes.

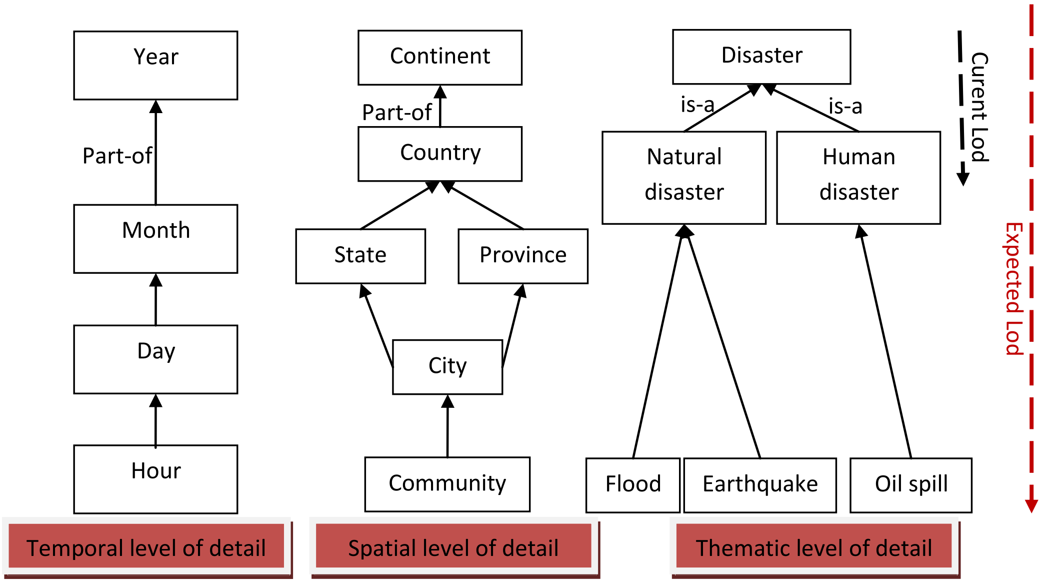


Figure 9.3 Example of hierarchy used to define the level of detail of input features

9.7.4 Measuring the Process Precision

The precision of a SIP indicates the degree of agreement between the level of detail of the input features and the level of detail taken into account by the process. When the level of detail taken into account by the process is equal to or higher than the level of detail of the input features, it means that the process is well adapted to the input and there is no information loss (with respect to the level of detail). Conversely, when the level of detail taken into account by the process is lower than the level of detail of the input features, the SIP creates semantic loss. For example, consider the global ontologies that are used by the semantic mapping process to support the identification of related concepts. If the most specific concepts defined in the global ontology (for example, "wetland") are more general than the input concepts being compared (for example, "emergent wetland" and "forested wetland"), the semantic mapping system will not be able to identify the input concepts. Therefore, the semantic mapping process' precision does not fit the input features'level of detail. Consider L_t^i, the level of detail of a single input feature i of type t, and L_t the level of detail of the SIP with respect to this type feature t. The precision of a process that takes as input a single feature of type t is measured by the following ratio:

$$\text{precision of process for a single input feature} = P_t(i) = \begin{cases} \dfrac{L_t^{'}}{L_t^{i}} & \text{if } L_t^{'} \leq L_t^{i} \\ 1 & else \end{cases}$$

Consider m, the number of input features for the given SIP. The global precision of the SIP is the sum of precision for each input feature, divided by the number of input features m:

$$\text{precision of semantic interoperability process } = \frac{1}{m}\left[P_t(1) + P_t(2) + \ldots + P_t(i) + \ldots P_t(m)\right]$$

The precision is maximal when L_t'>=L_t for all features, which means that the process always uses the finer level of precision possible. As the level of detail considered by the process becomes lower, precision goes down toward 0.

9.7.5 Measuring the Process Completeness

The completeness of a SIP indicates the degree of agreement between the features being present in the input and the current features that are taken into account by the SIP. The completeness of a SIP is therefore similar to precision. Let C be the number of elements used to define an input feature *i* (for example, for an input concept, the elements are the properties, the relations, the descriptors, etc.), and let C' be the number of these elements been used by the process. The completeness of a process that takes as input a single feature is measured by the following ratio:

$$\text{completeness of process for a single input feature} = \frac{C^{'}}{C(i)}$$

If we consider a number of m input features, completeness of a semantic interoperability process is defined as:

$$\text{completeness of process } = \frac{1}{m}\left[\frac{C'}{C(1)} + \frac{C'}{C(2)} + \ldots + \frac{C'}{C(m)}\right]$$

As for precision, a SIP preserves completeness when its value is 1 and preserves no completeness when its value is 0. Completeness and precision are two characteristics used to qualify a process in our framework. They are useful to indicate to the user whether the tools being used (coalition discovery tool, semantic mapping tool, etc.) are appropriate with respect to the input. Therefore, some tools being considered imprecise or incomplete for a given set of databases may be adequate for other databases whose semantics is defined with a lower level of detail or less features. Consequently, characteristics that measure quality or process can also be useful to select the appropriate matching or extraction tools.

9.7.6 Measuring the Consistency of Output

The consistency of output is a measure that applies only to SIP that issue set-based semantic relations. It is a relative measure of the number of conflicts that are created by the semantic relations between input features (such as concepts) issued by a SIP. Consider an input feature a_i related to another input feature b_j with the semantic mapping $m = (a_i, b_j, r)$, where r is the semantic relation between a_i and b_j. The consistency of this mapping can be assessed by verifying each Mapping Conflict Predicate provided in Bakillah et al. 2009. Consider $m = (a_1, b_1, r_{11})$, a mapping for which we want to compute the consistency with other mappings. Consider $V(a_1) = \{v_i^a\}$ and $V(b_1) = \{v_j^b\}$ the set of input features that are directly respectively related to a or b with a subsumption relation. These input features are related by a set of mappings M:

$$M = \{m_{ij} | m_{ij} = (v_i^a, v_j^b, r_{ij})\}$$

where each mapping m_{ij}, when compared to mapping m, can be consistent or inconsistent. Consider nc_i the set of consistent mappings that relate the input feature v_i^a of $V(a_1)$ to an input feature v_j^b of $V(b_1)$ and consider nn_i the set of inconsistent mappings that relate input feature v_i^a to an output feature v_j^b of $V(b_1)$. The consistency of mapping m is measured with:

$$\mathrm{C(m)} = \frac{1}{n}\sum_{i=1}^{n}\frac{nc_i}{nc_i + nn_i}$$

where n is the number of neighbour concepts of a_i : $n = \mathrm{card}(V(a_1))$. It is not necessary to consider the mappings from the point of view of neighbour concepts of b_j, since these mappings each have a reciprocal mapping established from the point of view of neighbours of a_i. In other words, if we determine the number of mappings inconsistent with mapping $m = (a_1, b_1, r_{11})$, we detect at the same time the semantic mappings that are inconsistent with the reciprocal mapping $m = (b_1, a_1, r_{11})$. The global consistency of output for a total number of mappings nm is given by:

$$Consistency\ of\ output = \frac{1}{nm}\sum_{k=1}^{nm} C(k)$$

Where C(k) is the consistency of mapping k.

9.7.7 Measuring the Explicitness of Output

The *explicitness of output* is analogous to the *explicitness of input*: it measures the ratio of the number of output features that are issued by a SIP to the number of expected features. For example, the expected features of a coalition's context issued by the coalition discovery process include at least one domain, one role, a geographical area and a time period. Let El^i_{expect} be the number of expected features (in the preceding example, $El^1_{\text{expect}} = 4$), and El^1_{present} the number of present features in a single output element i (for example, a single context i). The explicitness of this output element i is defined by the following ratio:

$$\text{explicitness of output for a single output element} = \frac{El^i_{present}}{El^i_{\text{expect}}}$$

If we consider an arbitrary number m of output elements (for example, the coalition discovery process issues m coalitions and their respective contexts), the global explicitness of output is given by their sum, normalized with m:

$$\text{explicitness of output} = \frac{1}{m}\left[\frac{El^1_{present}}{El^1_{\text{expect}}} + \frac{El^2_{present}}{El^2_{\text{expect}}} + \ldots + \frac{El^m_{present}}{El^m_{\text{expect}}}\right]$$

The number of expected features is set by the conceptual model for the output KR, but it can also be fine-tuned according to the user's requirements. The explicitness of the output is usefule as a relative measure: for example, in the case of semantic augmentation, we can compare the explicitness of the output with the explicitness of the input to verify if the semantic augmentation was effective.

9.7.8 Measuring Level of Detail of Output

The *level of detail of output* is analogous to the *level of detail of input*. It measures the level of detail of a SIP output feature, denoted $\text{LoD}^{i}_{\text{current}}$, with respect to the expected level of detail for this feature, denoted $\text{LoD}^{i}_{\text{expected}}$. The level of detail of each output feature is added for each of the *m* SIP output features:

$$\text{level of detail of output} = \frac{1}{m}\left[\frac{LoD^{1}_{current}}{LoD^{1}_{\exp ect}} + \frac{LoD^{2}_{current}}{LoD^{2}_{\exp ect}} + \ldots + \frac{LoD^{m}_{current}}{LoD^{m}_{\exp ect}}\right]$$

The level of detail and the explicitness of the output are necessarily the same as the level of detail and the explicitness of input, since any output element can become the input of another SIP.

9.7.9 Using Semantic Interoperability Quality Measures to Support Decision Making

The aim of this semantic interoperability quality framework was to provide a basis for the user to assess the quality of the semantic interoperability process. Although it is not in the scope of this thesis to develop an approach for the communication of the quality to the user, and thoroughly indicate how users can base their decision on the quality measures, we briefly indicate how the quality characteristics that we have developed can support the user to make sound decisions when different solutions are offered. For example, at the end of the semantic interoperability process, several concepts that match a query concept are proposed as solution to the user. How the quality characteristics that were presented above can support the identification of the optimal solution, especially given that in some cases they may be contradicting? We propose to use a multi-criteria analysis method to resolve this problem. Multi-criteria analysis refers to a category of methods that rely on several, sometimes contradicting criteria to select the best option or optimal solution to a problem.

More precisely, we propose to use the weight sum method, which aims at resolving complex problems where several alternative solutions exist, and several decision criteria with varying importance must be considered. The principle of the weight sum method is to build a decision matrix where possible solutions (lines of the matrix) are associated to values for each chosen decision criterion (columns of the matrix). In addition, a weight (percentage) is assigned to each criterion to indicate its relative importance. A score corresponding to the weighted sum of the criteria values is computed for each solution. The best solution is the one with maximal score. To apply this method to the quality of semantic interoperability, we have identified that:

- the different solutions to a query submitted by a user are the final results of the global semantic interoperability process, which are the different concepts retrieved by the semantic reconciliation component and which are proposed to answer the query (which could have been retrieved through different propagation strategies, different parameters for forming coalitions, etc.);
- the decision criteria are the characteristics (consistency of input, explicitness of input, … precision of process, etc.) for the quality of semantic interoperability;
- the weights for each characteristic are defined by the user.

The decision matrix is as follow:

	Quality characteristics (criteria)							
Semantic Interoperability Solutions	Consistency of input	Explicitness of input	Level of detail of input	Precision of process	Comprehensiveness of process	Consistency of output	Explicitness of output	Level of detail of output
Weights for quality characteristics	W_{ci}	W_{ei}	W_{li}	W_{pp}	W_{cp}	W_{co}	W_{eo}	W_{lo}
S1	V_{ci1}	V_{ei1}	V_{li1}	V_{pp1}	V_{cp1}	V_{co1}	V_{eo1}	V_{lo1}
S2	V_{ci2}	V_{ei2}	V_{li2}	V_{pp2}	V_{cp2}	V_{co2}	V_{eo2}	V_{lo2}
S3	V_{ci3}	V_{ei3}	V_{li3}	V_{pp3}	V_{cp3}	V_{co3}	V_{eo3}	V_{lo3}
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
Sn	V_{cin}	V_{ei4}	V_{lin}	V_{ppn}	V_{cpn}	V_{con}	V_{eon}	V_{lon}

The optimal solution is provided by the following:

$$\text{optimal semantic interoperability solution } = \max_i \sum_{j=1}^{N=6} w_j v_{ij}$$

Where j = {consistency of input, explicitness of input, level of detail of input, precision of process, comprehensiveness of process, consistency of output, explicitness of output, and level of detail of output}, i is the index referring to a semantic interoperability solution (i.e. a query result), and w_j is is the weight. Note that this approach is applicable for any of the semantic interoperability processes, not only at the very end of the process. Weights allow discarding the criteria that represent no interest to the user. To display the results to the user, we can use a diagram for each solution where the [0, 1] interval is partitioned into regions with qualitative denominations (Figure 9.4):

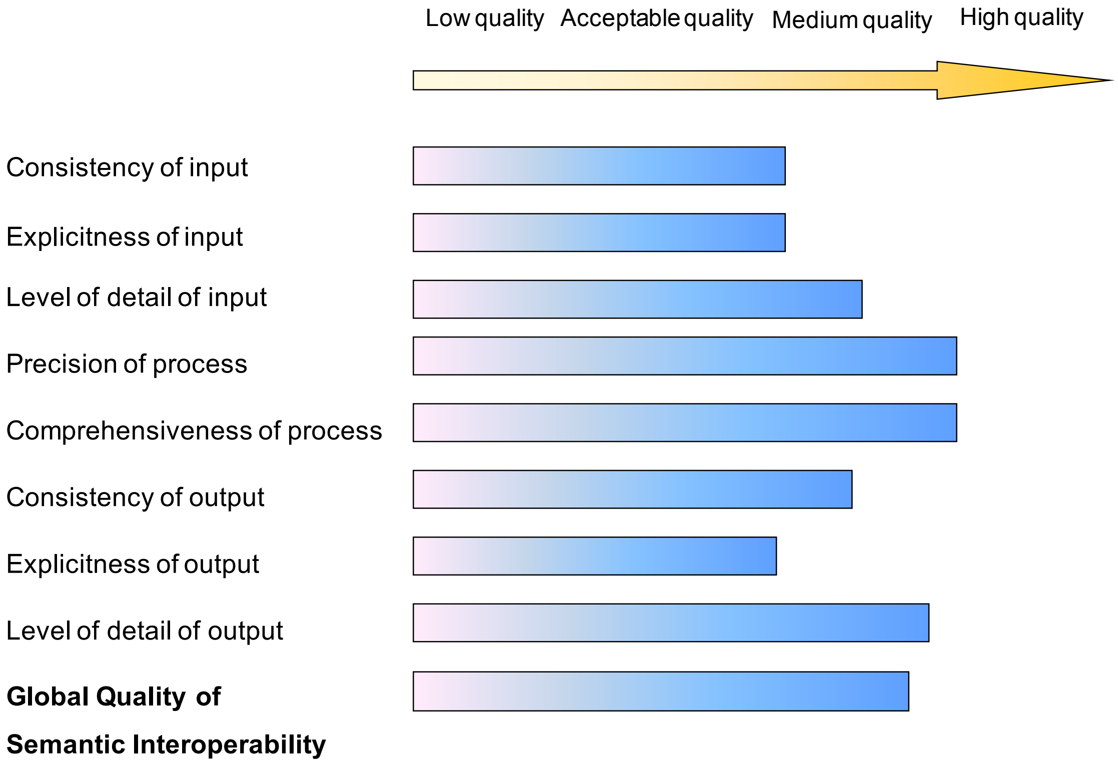


Figure 9.4 Ranges for Quality Characteristics

Besides supporting the user in decision-making, some quality characteristics can also be employed to test the proposed approach. In the next section, it will be shown how the measures are applied to further test the prototype.

9.8 Experimentation Method

In order to further establish the validity of the prototype that we have developed, we have assessed the performance and the behavior of the semantic interoperability processes with some of the quality measures presented above. In order to identify the most appropriate quality measures, we have considered that the goal of the experiment is to assess whether our approach improves semantic interoperability, either by (1) enriching semantics or (2) preserving and exploiting semantics that are being processed. Therefore, the characteristics for quality of input were not used because they do not assess these aspects. In addition, assessing some of the quality characteristics would have confirmed only what was more obvious or expected: for example, since the G-MAP was specifically designed to take as input MVAC ontologies, it was not necessary to assess the comprehensiveness of the G-MAP process. Therefore, in this experimentation, we have decided to assess the following aspects:

- precision of the coalition discovery process, and level of detail of the ouput (coalitions);
- explicitness gain produced by the MVAC semantic augmentation component;
- consistency of semantic mappings;
- precision and comprehensiveness of the query propagation component.

Table 9.2 summarizes and explains the choice of the evaluation criteria. The goal of each component is restated, and the evaluation measures were chosen to assess whether the goal was achieved. It must be noted that the ability of the G-MAP semantic reconciliation component to retrieve relevant concepts was tested in Chapter 3, and that the ability of the query propagation component to find relevant propagation paths was tested in Chapter 6.

Table 9.2 Choice for evaluation measures

Component	Goal of the component	Chosen evaluation measure
Coalition discovery component	Exploit the maximal level of detail in database descriptions to find groups of geospatial databases with meaningful descriptions	- Precision of process - Level of detail of output
Query propagation component	Exploit as much knowledge as possible on nodes of the network to find the propagation path that includes the most relevant databases with respect to a query	- Precision of process - Comprehensiveness of process[8]
MVAC component	Improve the explicitness of the input ontologies by adding features that will be exploited during semantic relconciliation process	- Explicitness gain
Semantic reconciliation component	Exploit as much semantics as possible to resolve semantic heterogeneities; issue semantic relations that are non-contradicting with existing semantic relations within ontologies.	- Consistency of output[9]

The details and results of the experimentations are provided in the following section.

9.9 Results of the Experimentations

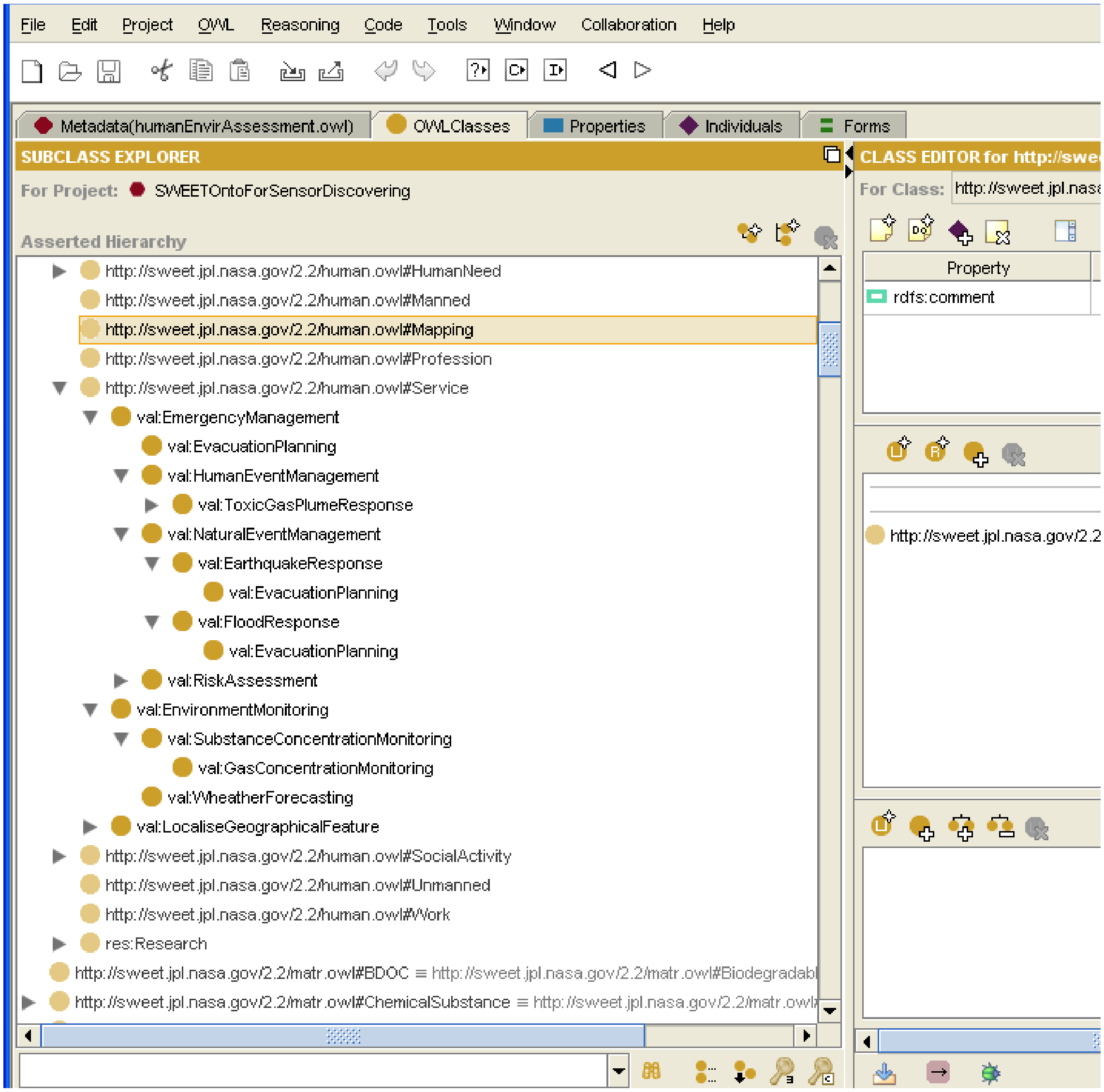
9.9.1 Experimentation of the Coalition Discovery Process

In the first experimentation stage, we tested the coalition discovery process (called CDP hereafter). The objectives of this experimentation stage are (1) to assess the ability of the CDP to take into account the LoD of input database contexts, and (2) the ability of the CDP to find meaningful coalitions, i.e. coalitions for which the context elements have a sufficient LoD to be considered as meaningful. Accordingly, to address the first objective, we measured the precision of the CDP according to the measure defined in Section 9.7.4. To address the second objective, we measured the LoD of the output of the CDP (i.e., the coalitions' contexts) with the measure defined in Section 9.7.8.

[8] Note that the query propagation component was also tested in Chapter 6.

[9] Note that the G-MAP semantic reconciliation component was also tested in Chapter 3.

<u>CDP precision assessment:</u> We compared the LoD of input context features (domain, role, geographical localization and temporal period) with the LoD that the CDP was able to process. To measure the LoD of spatial and temporal context features (geographical localization and temporal period), we used the hierarchy given in Figure 9.3. To measure the LoD of thematic features (function and domain), we built a hierarchy of functions and domains included in the sample of database contexts that were considered, based on the SWEET ontologies. An excerpt of this ontology is shown in Figure 9.5.



For Project: SWEETOntoForSensorDiscovering

For Class: http://sweet.jpl.nasa

```
▶ http://sweet.jpl.nasa.gov/2.2/human.owl#HumanNeed
  http://sweet.jpl.nasa.gov/2.2/human.owl#Manned
  http://sweet.jpl.nasa.gov/2.2/human.owl#Mapping
  http://sweet.jpl.nasa.gov/2.2/human.owl#Profession
▼ http://sweet.jpl.nasa.gov/2.2/human.owl#Service
    ▼ val:EmergencyManagement
          val:EvacuationPlanning
        ▼ val:HumanEventManagement
            ▶ val:ToxicGasPlumeResponse
        ▼ val:NaturalEventManagement
            ▼ val:EarthquakeResponse
                  val:EvacuationPlanning
            ▼ val:FloodResponse
                  val:EvacuationPlanning
        ▶ val:RiskAssessment
    ▼ val:EnvironmentMonitoring
        ▼ val:SubstanceConcentrationMonitoring
              val:GasConcentrationMonitoring
          val:WheatherForecasting
    ▶ val:LocaliseGeographicalFeature
▶ http://sweet.jpl.nasa.gov/2.2/human.owl#SocialActivity
  http://sweet.jpl.nasa.gov/2.2/human.owl#Unmanned
  http://sweet.jpl.nasa.gov/2.2/human.owl#Work
▶ res:Research
  http://sweet.jpl.nasa.gov/2.2/matr.owl#BDOC ≡ http://sweet.jpl.nasa.gov/2.2/matr.owl#Biodegradabl
▶ http://sweet.jpl.nasa.gov/2.2/matr.owl#ChemicalSubstance ≡ http://sweet.jpl.nasa.gov/2.2/matr.owl#
```

Property
rdfs:comment

http://sweet.jpl.nasa.gov/2.2

Figure 9.5 An excerpt of the ontology of domains and functions used for coalition discovery

The LoD of a domain or function corresponds to the level of this element of the hierarchy (as identified manually), whereas the LoD of the CDP corresponds to the LoD of the finest element that the CDP was able to identify and match with corresponding context features of other databases. We have done this assessment for nine database contexts, for which an example is presented in Annex 1.

The types of input context features were considered separately for a more detailed analysis of the CDP precision. Figure 9.6 presents the result of the evaluation.

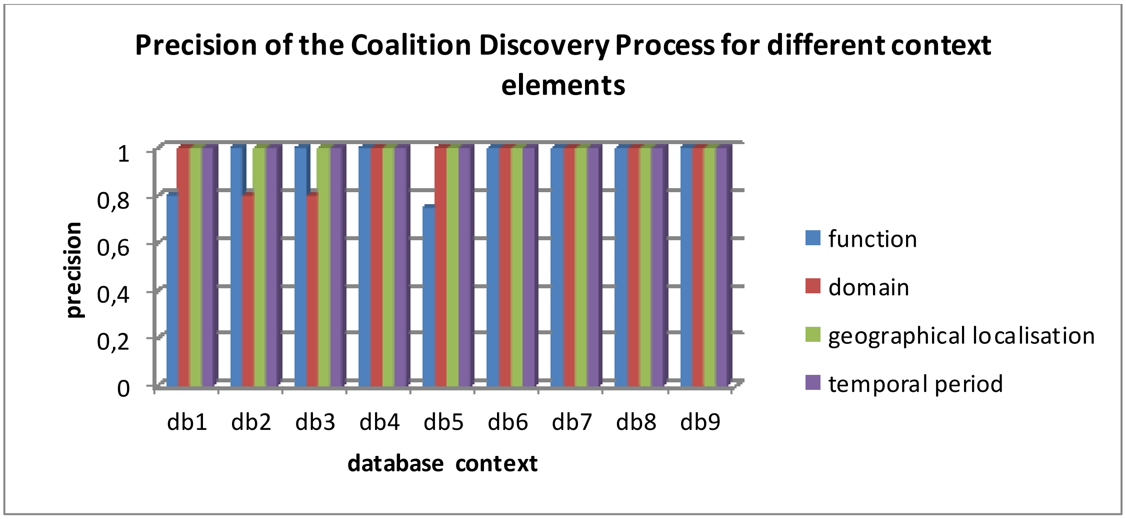


Figure 9.6 Assessment of the precision of the coalition discovering process

Concerning the precision of the CDP with respect to the geographical localisation, the precision was always 1 essentially because the LoD of input geographical localisations was somewhat low (consisting of countries or provinces). In our approach, the precision of the CDP with respect to the geographical localization depends mostly on the comprehensiveness of the ontology of places being used. However, it is possible that an input geographical localization does not exactly match the place name in the global ontology of places (e.g., ULaval vs Université Laval). A string-based matching approach can resolve most of the naming heterogeneities and improve the precision of the CDP with respect to the geographical localisation.

Concerning the precision of the CDP with respect to the temporal period, the precision was also 1 because the highest level of detail was the day, and matching temporal periods is straightforward.

However, the precision of the CDP with respect to the thematic context element (function and domain) is situated between 0,75 and 1. This is because in some instances, a function or domain could not be identified by the semantic attraction measurement process. For example, the function "identify material damage" of the Canadian Disaster Database (db5) could not be identified as a sub-function of "disaster response." The main issue being identified is the variety of functions that are mentioned in databases specifications, and the lack of an appropriate existing global ontology describing the various functions for geospatial databases, which would support the CDP in discovering sub-functions. However, we note that it is out of the scope of this thesis to develop appropriate global ontologies of functions and domains, although our approach relies on such ontologies. The experimentation nevertheless indicates that in the majority of cases, the precision of the CDP was high.

Level of detail of CDP output: The output of the CDP is a set of coalitions with their respective context. The set of coalitions being issued depends on the semantic attraction threshold selected by the user (see Figure 8.8). For example, if the user selects a low semantic attraction threshold (such as 0,1), the resulting coalitions are very inclusive (they contain more databases, because the condition to be part of the coalition is not very restrictive). As a result, when the context of the coalition is computed by merging contexts of member databases, the context is usually very broad and with very low level of detail. Consequently, we have made the hypothesis that the level of detail of the coalition contexts will increase with the semantic attraction threshold.

An excerpt of the coalition contexts that were formed is provided in Annex 2. We have assessed the LoD of the computed coalition contexts for different values of the semantic attraction threshold, ranging between 0,05 and 0,60 (Figure 9.7). The range was restricted to 0,60 because above that threshold, no coalitions could be formed.

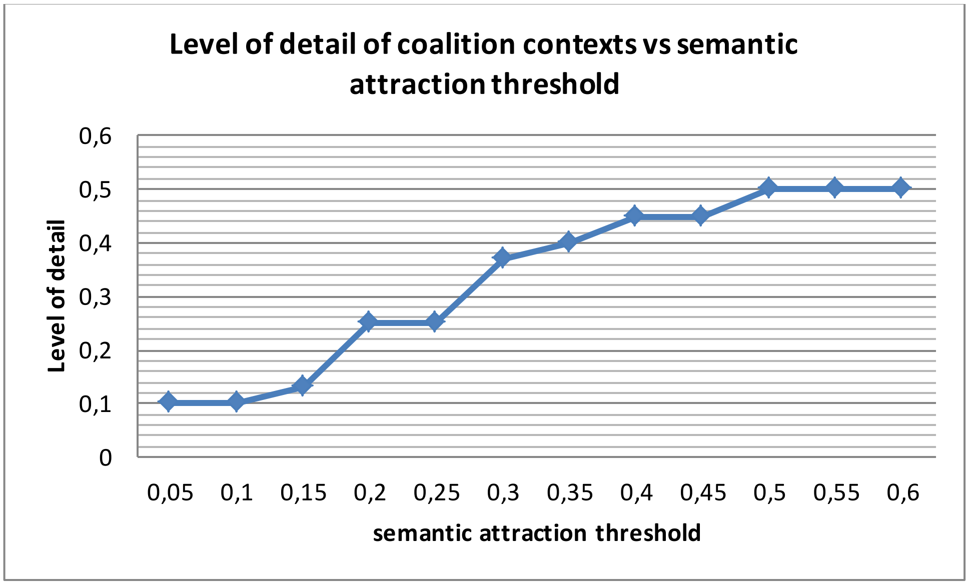


Figure 9.7 Assessment of the level of detail of the coalition discovering process output

The experiment confirmed the hypothesis, i.e. below the 0,25 semantic attraction threshold, the level of detail is somewhat low, with for instance coalitions whose context is defined as {domain= hydrography, role = disaster management, geographical location = Canada, and temporal period = 1900-2011}. Such coalitions, because they include a lot of databases that have little similarity are too broad to be meaningful. Above the 0,30 threshold, more meaningful coalitions have a greater level of detail, for example, {domain= road network; role = evacuation planning, identify road classification; geographical location = Canada; and temporal period = 2011}. The experiment shows that when an appropriate semantic attraction threshold is chosen, meaningful coalitions with significative level of detail can be discovered. The user can understand at a glance the nature of data held by the members of the coalition and the coalition contexts can be reused efficiently during query propagation.

9.9.2 Experimentation of the MVAC Augmentation process

In the second experimentation stage, we tested the MVAC augmentation process. The objective of this experiment was to assess the ability of the MVAC augmentation component to improve the expliciteness of input ontologies. To do so, we measured the explicitness gain produced by the MVAC augmentation component. The explicitness gain is defined as:

$$\text{Explicitness gain} = \text{explicitness of the output} - \text{explicitness of the input}$$

To measure the explicitness of the input and of the output, we set the types of expected features as follow: {property, relations, spatial descriptor, temporal descriptor, context, dependency} (i.e., m = 6), and the number of expected features El_{expect} to the maximum number of this type of features among all the augmented concepts. The following table shows the El_{expect} values for each type of feature:

Type of feature	El_{expect}
property	9
relations	5
spatial descriptor	4
temporal descriptor	2
context	3
dependency	4

For example, the maximum number of dependencies that were extracted for a single concept within the sample is 4, and the maximum of contexts that were extracted for a single concept is 3. As an example, the following is the concept "road" from the BDTQ:

```
<concept>
<nomconcept>road</nomconcept>
<def>infrastructure for the displacement of vehicles. Includes roads in
construction and abandoned roads</def>
            <proprietes>
                <ensproprietesthem>
                  <proprietethem>
                        <nom>surface</nom>
                        <ensval>
                            <val>paved</val>
                            <val>unpaved</val>
                        </ensval>
                    </proprietethem>
                    <proprietethem>
                        <nom>status</nom>
                        <ensval>
                            <val>praticable</val>
                            <val>not praticable</val>
                        </ensval>
                    </proprietethem>
                    <proprietethem>
                        <nom>road classification</nom>
                        <ensval>
                            <val>highway</val>
                            <val>street</val>
                            <val>rural road</val>
                            <val>access road</val>
                        </ensval>
                    </proprietethem>
                </ensproprietesthem>
                <ensproprietesspat>
                    <proprietespat>
                        <nom>geometry</nom>
                        <ensval>
                            <val>line</val>
                            <val>polygon</val>
                        </ensval>
                    </proprietespat>
                </ensproprietesspat>
            </proprietes>
            <relations>
                <ensrelthem>
                  <relthem>
                        <nom>is-a</nom>
                        <memb1>road</memb1>
                        <memb2>road network feature</memb2>
                    </relthem>
                </ensrelthem>
                <ensrelspat>
                  <relspat>
                        <nom>overlap</nom>
                         <memb1>road</memb1>
                         <memb2>vegetation</memb2>
                         <memb2>bridge</memb2>
                  </relspat>
                  <relspat>
                      <nom>in</nom>
                         <memb1>road</memb1>
                         <memb2>rural area</memb2>
                         <memb2>urban area</memb2>
                   </relspat>
                </ensrelspat>
            </relations>
            <descripteurs>
                  <ensdescrspat>
                        <descrspat>
                              <nom>geometry represents</nom>
                        <ensval>
                            <val>central axis of roadway</val>
                            <val>roadway</val>
                        </ensval>
                  </descrspat>
                  <descrspat>
                        <nom>width</nom>
                        <ensval>
                            <val>less than 20m</val>
                            <val>20m and over</val>
                        </ensval>
                        </descrspat>
                        <descrspat>
                              <nom>shape</nom>
                        <ensval>
                            <val>curve</val>
                        </ensval>
                        </descrspat>
                  </ensdescrspat>
            </descripteurs>
            </concept>
```

Since this concept "road" has 4 properties, 3 relations with other concepts and 3 spatial descriptors, but no other types of features, its relative explicitness is 1/6*(4/9 + 3/5 + 3/4) = 0.30. The augmented concept "road" (provided in Annex 4) contains, in addition to the features of the original concepts, 3 dependencies and 3 contexts. As a result, its relative explicitness is 1/6*(4/9 + 3/5 + 3/4+ 3/3 + 3/4) = 0.59. Therefore, the explicitness gain for this concept is 0.59 - 0.30 = 0.29. We have augmented nine ontology samples and performed the same calculations. Figure 9.8 presents the results of the experimentation.

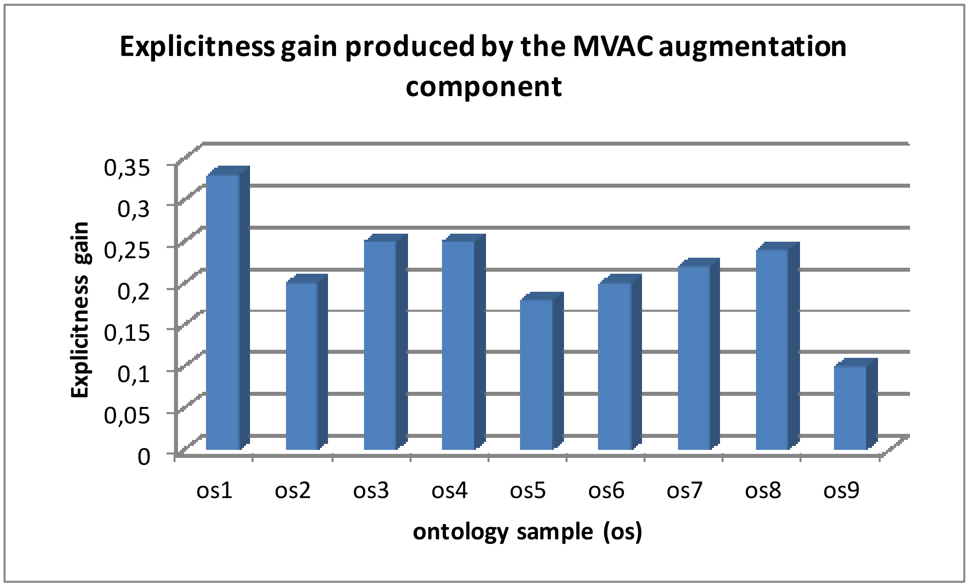


Figure 9.8 Explicitness gain produced by the MVAC augmentation component

The overall average expressiveness gain is of about 0,22, meaning that the overall performance of the MVAC is significative. It must be noted that the support and confidence threshold for selecting dependencies were set both to 0,80. During the testing, it was noted that a lower threshold selects dependencies that are not significant, while a higher threshold discards dependencies that are considered valid from the perspective of the application domain. Some concepts could not be augmented because no dependency exists. Similarly, some concepts had a poor definition. For example, in the BDTQ, a tunnel is described as "underground tunnel for the passage of a road under a river, an arm of the sea or across a land elevation." But the water canal is only defined as an "artificial waterway," and it was not possible to extract different contexts and therefore different views for this concept and other concepts with poor definition. Consequently, the performance of the MVAC is also limited by the richness of input ontologies. This is because the role of the MVAC is to enrich the concept by making explicit what is implicit, but without using external resources. The semantic enrichment with external resources could be an avenue to further improve the MVAC component, but several issues would have to be addressed, such as the trust given to external resources, and selecting the appropriate resources.

9.9.3 Experimentation of the Semantic Reconciliation Component

In the third step of the experimentation, we tested the semantic reconciliation process. In chapter 3, the accuracy and recall of G-MAP were already tested to demonstrate the positive impact of the augmentation. Also, because G-MAP was designed to take as input the MVAC ontologies, testing the precision and the comprehensiveness of the G-MAP process is not required. In this section, the objective was to assess the ability of G-MAP to detect consistent semantic mappings, i.e. semantic mappings that do not create conflicts with existing relations within ontologies. To do so, we assessed the consistency of a sample of semantic mappings between original "non-augmented" ontologies, and compared the results with the consistency of a sample of semantic mappings between the same ontology samples that were now "augmented." The experimentation was conducted for 18 pairs of ontology samples. The consistency was assessed according to the definition provided in Section 9.7.6 and the Mapping Conflict Predicates (MCP). Consistency is measured according to the number of mappings that verify consistency predicates. For example, we have computed the following semantic mappings between concepts of a pair of sample ontologies (where the table reads as "hydroelectric feature" included in "lake," for example):

		BDTA	
		Hydrographic feature	lake
EC meteorological database	Hydrographic feature	m1: includes	m2: included in
	Hydroelectric reservoir	m3: overlaps	m4: overlaps

The mapping m1 is inconsistent with mapping m2, because it verifies the mapping conflict predicate MCP 3. To compute the consistency of mapping m1, for example, we consider the number of mappings that are inconsistent with m1 (i.e., mapping that verify MCPs) in the neigborhood of m1 (as defined in Section 9.7.6) (in this case, m2) and the number of mappings that are consistent with m1 (in this case, m3 and m4). Using the formula provided in Section 9.6.7, with n = 2 the number of concepts in the neighbourhood of the source concept of m1, we obtain the consistency of m1: ½*(2/3+2/3) = 2/3. We performed this calculation for a sample of mappings and obtained the results displayed in Figure 9.9.

The result shows that the consistency is preserved despites augmentation. This is because the dependencies that were discovered generally do not create conflict with the hierarchy of concepts. In only one case (op4), the introduction of augmented concepts resulted in an inconsistent mapping which reduced the consistency to 0,80. This indicates that very few dependencies induce false mappings, with respect to semantic mappings without augmented features. The conclusion of the experimentation is, however, that the consistency measure can be used as an additional tool to help to select valid dependencies and reject invalid ones. Nevertheless, the consistency of mappings can be used as a warning for the user to indicate potentially misleading mappings.

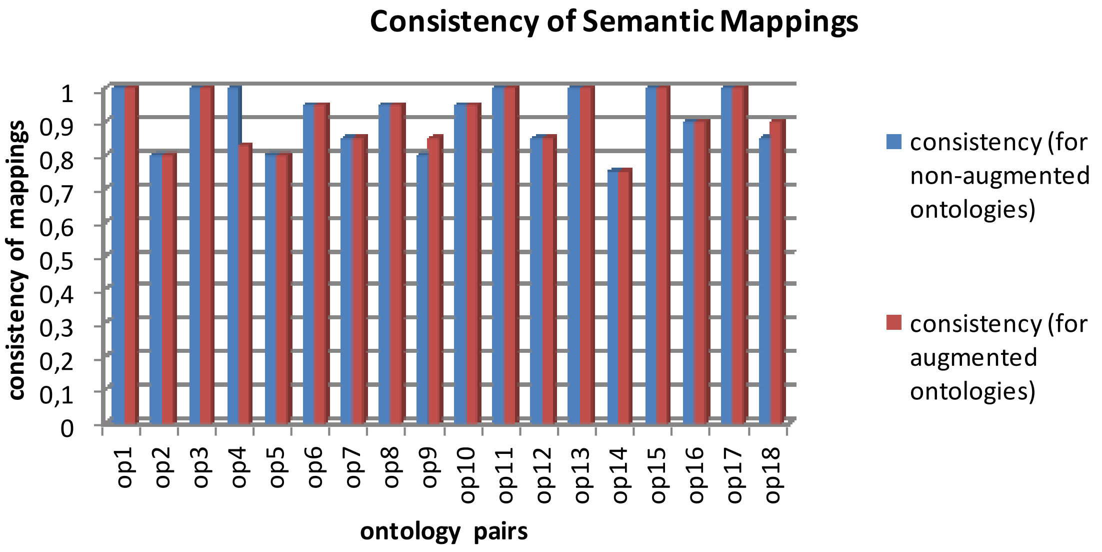


Figure 9.9 Consistency of semantic mappings with and without augmentation

9.9.4 Experimentation of the Query Propagation Component

In the last experimentation phase, we conducted additional experimentations on the query propagation component (performance experimentations were already performed and presented in Chapter 6). The objective of the experiment was to assess the ability of the query propagation approach to identify the sources that are relevant to the query. To do so, the query propagation component should be able to consider all elements of the input with the adequate level of precision and comprehensiveness. Therefore, we assessed the level of precision and comprehensiveness for the three different strategies (coalition-based, database context-based, and memory-based strategies). This experiment is similar to the one perfomed on coalition mining, and the results, shown in Figure 9.10 to 9.12, are similar too. The level of detail (used to measure the precision) and expected number of input features (used to measure the comprehensiveness) were set in the same manner as in the experimentation of the coalition discovery process in Section 9.9.1. The average comprehensiveness and precision for all sources is shown for each strategy, and for a set of ten queries.

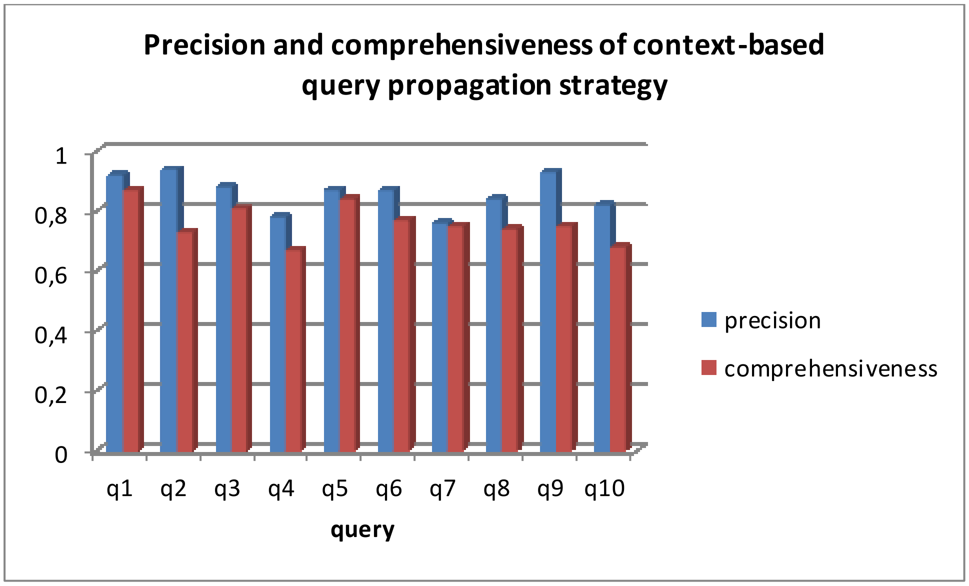


Figure 9.10 Precision and comprehensiveness of the context-based query propagation strategy for a set of 10 queries

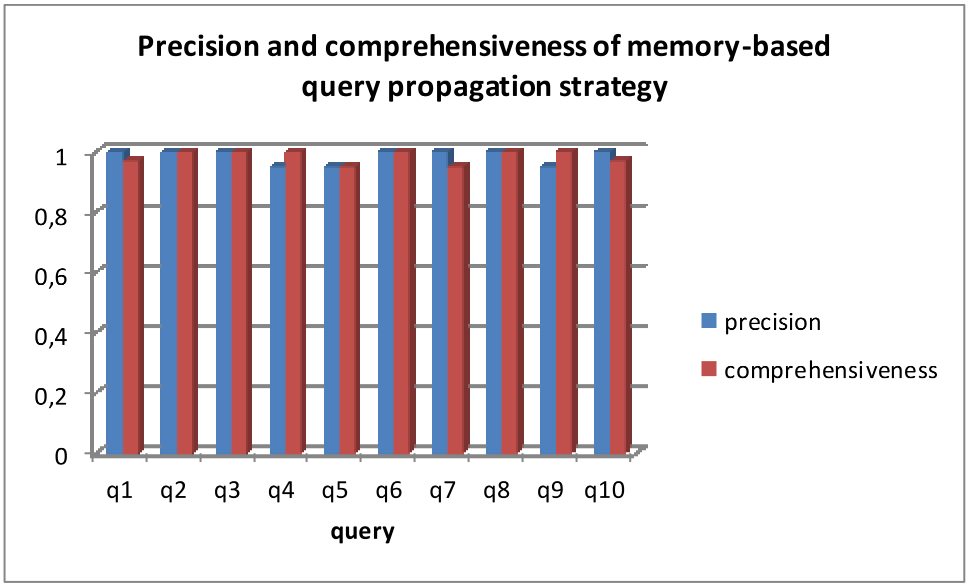


Figure 9.11 Precision and comprehensiveness of the memory-based query propagation strategy for a set of 10 queries

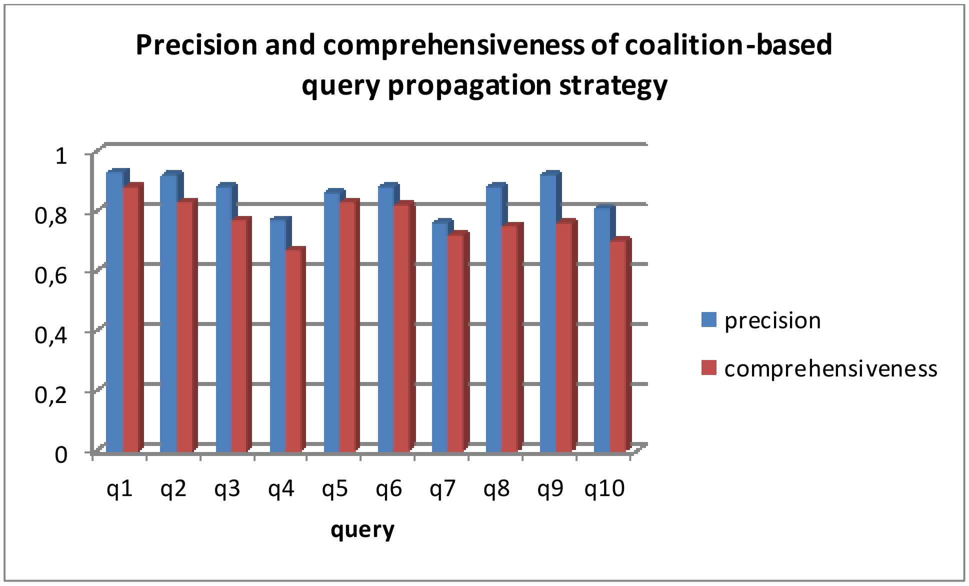


Figure 9.12 Precision and comprehensiveness of the coalition-based query propagation strategy for a set of 10 queries

While the precision and comprehensiveness of query propagation processes are generally high, the performance of the memory-based strategy is higher. This is due to the fact that in this strategy, current query is compared to the past queries, and queries are less complex (composed of a smaller number of elements) than context descriptions, which are used in the two other strategies. For example, a context of a database can contain a set of several domains and functions, while a query contains only one function and one domain. The coalition-based and context-based strategies follow the same pattern since both are based on the comparison of databases'contexts or coalition's contexts, which have the same structure. The limitations of the query propagation approach, with respect to the identification of relevant database, is also related, as for the coalition discovery approach, to the richness and appropriate character of the global ontologies being used.

The experimentation shows that the proposed semantic interoperability quality framework can help to understand the ability of a semantic interoperability process to enrich semantics, preserve existing semantics and avoid semantic loss. From the point of view of the user, the benefit of this framework is therefore to support the gradual assessement of the global semantic interoperability process, which empowers him or her to make sound decision when discovering and sharing geospatial data.

9.10 Discussion and Conclusion

In this chapter, we have presented a framework for the quality of semantic interoperability, based on the framework of semantic mapping quality we have previously developed. This framework includes the definition of quality characteristics and their measures. In order to demonstrate that our approach improves real time semantic interoperability in ad hoc networks of geospatial databases, we have used some of these quality measures to conduct experimentations on the components of the approach. The experimentations showed that the approach can improve semantic explicitness (with the MVAC augmentation component), and that the semantic interoperability processes which support semantic reconciliation, discovery and communication at various levels preserve and exploit the richness of input semantics.

This experimentation also raises issues that could be considered in future research to further enrich the approach. We have indicated that the ability of the coalition discovery approach and the query propagation approach to find relevant sources is partly dependent on the suitability of the global ontologies being used with respect to the input context elements. In particular, a global ontology of roles for geospatial databases would support enhanced coalition formation, as it appears that the role is very important to make the coalitions meaningful. We have also indicated that in some instances, the ability of the MVAC semantic augmentation component to improve explicitness is limited by the constraint of using only available knowledge, and not knowledge that could come from external sources. Further research could investigate what other reliable sources of external knowledge could be exploited to enrich existing ontologies. Although it was not an objective of this thesis, the communication of the quality of semantic interoperability to the user is also an important factor. For example, the quality characteristics could be used as indicators that can be communicated to the user as warnings at different points of the semantic interoperability process. This could improve support to decision-making during the semantic interoperability process.

CHAPTER 10

Conclusion

10.1 Summary

This thesis addressed the problem of real time semantic interoperability in ad hoc networks of geospatial databases.

In Chapter 1, the context of the research, the general and specific problems and the objectives of the thesis were defined. It was stated that this research aims at proposing a solution for real time semantic interoperability in ad hoc networks of geospatial databases, which is a very complex issue because of the different aspects involved. Indeed, we noted that in order to achieve real time semantic interoperability in ad hoc networks of geospatial databases, we had to go beyond the problem of resolving semantic heterogeneity, as we also need to consider the large number of data sources, which has required a solution to organize and search through the network, and the dynamicity of the network. This means that we had to address several specific problems: the discovering of groups of geospatial databases that can form meaningful coalitions, the problem of propagating geospatial queries to relevant databases of the network, the issue of poor knowledge representation, and more particularly, the representation of geospatial concepts, and the problem of semantic reconciliation of heterogeneous ontologies of databases.

In Chapter 2, background and state of the art concerning topics related to semantic interoperability was presented. First, a review on ad hoc networks and their role in the geospatial domain as well as their role to support interoperability was given. Then, we presented the problems related to semantic interoperability, including the fundamental notions of heterogeneity, implicitness, and the meaning of real time. A background on ontologies was provided, comprising an explanation of their role with respect to semantic interoperability and ontology languages. Finally, a state of the art on core issues, including discovery and formation of groups in networks, knowledge representation and extraction, semantic mapping and semantic similarity, as well as query propagation, was given. The literature review revealed that existing approaches are not fully adapted to ad hoc networks of geospatial databases.

In Chapter 3, a conceptual framework for real time semantic interoperability of geospatial databases was proposed. We introduced the fundamentals of social networks as a basis of our framework. The characteristics of social networks were employed to justify the need of the components that participate in the proposed framework. The framework includes a conceptual model for geospatial databases coalitions, whose semantics are defined by their context, and represents the kind of coalitions that members of social networks can form according to different common criteria. To address the complex issue of geospatial concept representation, a conceptual model for the Multi-View Augmented Concept (MVAC) was proposed to include semantics of spatiotemporal properties, views of the concept that are valid in different contexts, and dependencies between features of the concept. It is important to note that following the development of this augmented concept model, whose complexity was meant to improve the semantic richness, we came to the conclusion that existing semantic mapping approaches were not adequate to support comparison of these augmented and multi-view concepts. The G-MAP gradual and augmented semantic mapping approach was developed specifically to address this issue and to support resolution of semantic heterogeneity problems on the thematic, spatial and temporal dimensions. G-MAP was specifically tailored to the MVAC Model. Finally, real time query propagation strategies that reproduce diverse abilities of social network members to communicate and disseminate information were outlined.

Chapters 4 to 7 correspond to several published or submitted articles that proposed solutions to the specific problems of this thesis and develop the approaches corresponding to the conceptual models for each component proposed in the conceptual framework developed in Chapter 3.

In Chapter 4, an approach based on network analysis and semantic attraction was proposed for the discovery of geospatial databases coalitions. To the best of our knowledge, our approach was unique, since it introduced social network analysis in support of semantic interoperability. In this approach, nodes (representing databases and their users) that act as central attractors are detected and coalitions are formed around them. Our idea was to reproduce the behaviour of social network members who gather around central leaders. Experiments showed that this idea was applicable and support the discovery of meaningful coalitions of databases.

In Chapter 5, the approach for the MVAC Semantic augmentation tool was proposed in an article published in the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science, Hong Kong, 26-28 May 2010. The MVAC semantic augmentation tool implements context extraction (which was only partly automated), view extraction and semantic augmentation based on rule mining techniques. The integrated MVAC Tool allows generating MVAC concepts from original concepts taken in ontologies. Besides enriching semantics, the tool introduced the idea of multi-view, context-dependent semantic interoperability. The MVAC approach and tool was also validated in a paper published in the *Journal of Earth Science and Engineering*.

In Chapter 6, an approach for real time query propagation was proposed. It introduced three complementary strategies: a first strategy uses existing coalitions of the network; the second strategy determines the propagation path according to context affinity between the query and the databases; the third strategy uses the knowledge about queries that were previously answered to forward queries to relevant databases. The strategies are formalized with the Lightweight Coordination Calculus (LCC), an adequate framework to simulate coordination in a dynamic setting according to social constraints. One of the original aspects of our approach was to propose different and complementary ways of selecting query reciepients through these strategies, instead of relying on a single type of criterion, as existing query propagation approaches do. Also, during the development of the query propagation approach, we noted that changes in the network, including the addition or removal of a source, were not taken into account by the existing query propagation approaches. As an additional contribution, an algorithm was proposed to simulate the changes in the strategies when a change in the network occurs. The comparative experiments show that strategies are complementary since they perform differently in terms of accuracy, recall and scalability.

In Chapter 7, the article entitled "SIM-NET: A View-Based Semantic Similarity Model for Ad Hoc Network of Geospatial Databases" that was published in 2009 in *Transaction in GIS* is proposed. This article indicates that no existing semantic similarity models were specifically designated for ad hoc networks and it proposes a new model based on Description Logics (DL) to fulfil this need. The characteristics of this model are its ability to take into account concepts that describe different or similar domains, the influence of neighbours in the network, and it takes as input concepts with several views, which makes it adapted to the MVAC model. In addition, an algorithm was proposed to simulate the behaviour of SIM-NET in a dynamic environment.

In Chapter 8, the implementation of the whole approach with a prototype was presented. The chapter starts by presenting the architecture and technologies used, and validates the global approach with a case study. The case study demonstrated the feasibility of the approach and the usefulness of its features for the user. It also showed that the different components of the approach could be successfully integrated to achieve the ultimate goal, real time semantic interoperability in ad hoc networks of geospatial databases.

Finally, Chapter 9 made further contributions by introducing the new concept of quality of semantic interoperability, for which a framework was suggested. The conceptual framework for quality of semantic interoperability extends a previous framework presented in "Elements of Semantic Mapping Quality: A Theoretical Framework", which was published in *Quality Aspects in Spatial Data Mining*. The framework for quality of semantic interoperability was applied to conduct further experimentations on the prototype and demonstrate that the proposed semantic interoperability quality framework can help to understand the ability of a semantic interoperability process to enrich semantics, preserve existing semantics and avoid semantic loss.

10.2 Discussion

In this thesis, we have presented an approach for real time semantic interoperability that was designed to support meaningful geospatial data sharing among participants of a dynamic network, geospatial data reuse, and sound interpretation of exchanged data. From the user's perspective, this approach provides several advantages. Among them, we mention that the context of geospatial databases can be expressed with different types of constraints on thematic, spatial and temporal features. This approach is more expressive than some other group formation approaches, for example those of Khambatti et al. (2002), Mika (2004) and Lumineau and Doucet (2004), which were not dedicated to the geospatial domain, and where user's need, interests or domain of expertise are expressed as keywords. In fact, the problem of group discovery in networks was very rarely addressed in the geospatial domain. In the proposed approach, any database can also be part of several coalitions at the same time, so its users can manage different tasks requiring collaboration with others in parallel. In this thesis, while we have modeled coalition requirements as constraints and proposed an algorithm for the discovering of coalitions, we did not address more complex cases where some members do not agree, which would necessitate the development of a protocol supporting negotiation among members of the network. While examples of such negotiation protocols were already proposed, especially for Multi-Agent Systems (MAS) (ex: Dang et al. 2003, Sauro 2005, Oravec et al. 2007, Zheng et al. 2008, Boella et al. 2009, Van der Torre and Villata, 2009), it would be interesting to explore how such negotiation approaches could benefit from being “semantically-enabled” and how they can be adapted to our constraint-based definition of coalition requirement, including geospatial aspects.

With the MVAC tool, the user is also able to make the semantics of his or her data more expressive. The different views of the concept allow representing the latter from different points of view, depending on a given context. Not only different users can compare the meaning of the concepts defined in their ontologies, they can as well compare their contexts, with the G-MAP semantic mapping tool that is multi-view oriented. While the multi-view approach is not new (the concept was explored both in the geospatial database realm and ontologies, see Chapter 5), the difference of our approach is to automatically extract the views based on context rules. In addition, it would be interesting to investigate how the the different kinds of contexts (functional, spatial, temporal, etc.) would allow different ways of extracting views. Meanwhile, we also mentioned that the automatic extraction of contexts remains an issue which is complex and goes beyong the scope of this thesis, as more extensive research on natural language processing would be required.

In Chapter 3, it was also demonstrated that using the MVAC representation improves the performance of the semantic mapping process; from the implementation point of view, it would be interesting, however, to provide the user with a more interactive way of managing semantic mappings that were augmented. The G-MAP semantic mapping tool is not only useful to the user who would like to retrieve, within ontologies describing other databases of the network, the concepts that correspond to its query; it is also essential to identify semantic conflicts between different conceptual representations. Undetected conflicts lead to misinterpretation of data being exchanged, and therefore may lead to false decisions. The G-MAP identifies and highlights the differences of meaning between thematic and especially spatiotemporal aspects of concept that may lead to misinterpretation of geospatial data. It is a fundamental tool for geospatial data reuse.

The multi-strategy, real-time propagation approach is a tool that supports the users in finding which sources can answer their queries and therefore which sources must be semantically reconciled to answer the query. The different strategies are effective at different scales. A user may choose to ask the question only inside one of the coalitions that he or she is member of; this strategy is useful when the user does not want to query a large number of sources and he or she has some knowledge of the sources inside the coalition that allow him or her to expect that those sources will provide a satisfying answer. In longlasting coalitions, where some tasks may be performed repeatedly, it is also expected that the memory-based strategy will perform better, since each node has the possibility to store a greater number or queries in its memory. We also think that incorporating a learning algorithm to the memory-based strategy could improve its performance, but this would require extensive studies and substantive additional research to adapt such algorithms to our approach. The coalition-based strategy, where a query is forwarded to coalitions having similar contexts, covers a different kind of requirement, where the query cannot necessarily be answered by members of a single coalition, or the user need information coming from more diverse sources. This strategy will also support the user whose query concerns several domains and there is no existing coalition at this time to answer the query. From a conceptual perspective, developing various strategies bring more flexibility into the query propagation approach.

The main objective of this thesis was "to propose a framework and an approach for real time semantic interoperability in ad hoc networks of geospatial databases." The framework proposed in Chapter 3 and the subsequent developments in the following chapters show that the main objective was achieved, as well as the specific objectives. They also demonstrate that the hypothesis, which stated that "real time semantic interoperability in ad hoc networks of geospatial databases can be achieved with an approach that integrates four main components, that is, discovering coalitions of geospatial databases, real time query propagation, semantic augmentation of geospatial concepts, and semantic reconciliation," is plausible and correct. A global validation of the approach was made by a demonstration of the prototype.

10.3 General Conclusions

The following conclusions can be drawn:

- It is possible to combine solutions that address some issues of semantic interoperability of geospatial data and semantic interoperability in ad hoc networks to provide an original solution. Until now, existing approaches are either targeted as geospatial issues or for networks of non-geospatial sources from the larger computer science domain;

- One of the main bottlenecks of semantic interoperability frameworks is the richness of the semantics. Even in the geospatial domain, the semantics of spatial and temporal properties are sometimes poorly represented or assumed to be as any other concept, and mostly the “context”-related issues are often disregarded. A better knowledge representation model for “concepts” can improve semantic interoperability; however, it is somewhat useless if a correspondingly semantic mapping model is not available to take as input this representation.

-The new MVAC Model can be used a starting point to further develop the concept of context-oriented semantic interoperability, where the solution proposed to the user depends on the context he or she used. Visualization aspects concerning available contexts could be further explored and improved to enhance this concept. In the same way, a multi-level semantic interoperability process, where the user can choose between different levels of definition of the concept (in this case, concept, multi-view concept and augmented multi-view concept) is a new idea that was uncovered by the MVAC. We argue that many other improvements to the knowledge representation model are possible, that will enlarge the flexibility of semantic interoperability solutions that are more adaptable to the user’s context, knowledge, and requirements.

-The overall approach requires a certain amount of preliminary work by an expert who should have a sound knowledge of the domain. For example, context rules and database context descriptions have to be generated by experts. Additional automated context extraction methods and annotation methods, for example based on linguistic analysis method, could improve the approach.

10.4 Research Perspectives and Future Work

The approach presented in this thesis has created several new research perspectives that could enhance the sharing of geospatial data within an ad hoc network. Possible future research avenues include the following:

- The coupling of the coalition mining algorithm with tools for geo-collaboration in GIS would be interesting to investigate and could greatly enhance the latter. Geo-collaboration tools for GIS usually propose platform and user interfaces to support collaborative work, but do not include a method to discover who can collaborate and form groups. However, the issue of discovering groups of potential stakeholders becomes significant when considering a large number of potential participants and sources, and unexpected scenarios, such as in disaster response. The increasing pervasiveness of wireless data collection and generation devices, which enable the formation of ad hoc networks, as well as the emergent tools that enable users of geospatial data to become producers of volunteered geographic information, is likely to increase the need for such platforms.
- The other potential usages of the Multi-View Augmented Concept model are multiple. We expect that the MVAC model can be useful to support semantic interoperability of geospatial web services. For this, the MVAC can be used to enhance the description of web geo-services. The MVAC could be the basis of an approach for semantic mapping and composition of web geo-services. Another possible research perspective is the utilisation of the MVAC to describe the semantics of 3D models, and therefore support the integration of different 3D models, such as city models. More specifically, the reasoning capability of the MVAC can be adapted and enhanced to infer the semantics of the different spatial dimensions of the model.
- The different strategies for query propagation offer different ways to select query recipients; however, we believe that the query propagation strategies could be further enhanced with learning algorithms that learn “what is the best strategy”. In addition, a learning algorithm can be useful to enhance the memory-based strategy, that is, to learn similar queries.
- In this thesis, simple queries were composed of a concept only. However, queries can be more complex and involve any kind of concept feature and constraints. In future work, it would be interesting to further improve the query propagation strategies by considering various kinds of complex queries.

-An interesting avenue would be to employ the approach within dynamic geo-sensor networks that monitor environmental conditions. For instance, the coalition mining algorithm could be adapted to find clusters of sensors that detect or monitor similar properties or events (from the semantic of the primary events or state that they detect), and infer larger-scale events. This approach, because it is semantic-based, would complement the sensor approaches that focus solely on spatial location of sensors. In addition, we are currently investigating how the query propagation approach can be adapted to the propagation of queries in sensor networks. Sensor networks are contrained by specific types of topologies, and it would be interesting to investigate how the strategies could be adapted to these topologies.

-The Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) initiative enables the access to sensor data over the Web. The vast amount of available sensors gathering data on various environmental phenomena requires the development of sensor data discovery mechanisms that enable the users to find sensor data relevant to their needs. To be efficient, sensor data discovery mechanisms must rely on explicit semantic models of sensor data and address the problem of semantic heterogeneity specific to sensor data. We also believe that several elements of our approach could improve semantic enablement of the Sensor Web. More specifically, the MVAC model could be used as a basis to enhance the richness of the SensorML standard, which is currently used to describe sensor data and observations on the Sensor Web. Further work could then be done to adapt the semantic reconciliation tools that we have developed in this thesis to this enriched SensorML model.

-Finally, we have proposed the basis of a framework for quality of semantic interoperability. This is a very new concept, which opens many research opportunities, given that to the best of our knowledge, there exists no comprehensive approach to assess the quality of a semantic interoperability solution. We argue that the use of a set of semantic interoperability quality metrics and assessment tool is strongly needed to support the development of appropriate semantically interoperable systems. More specifically, the issue of communicating quality of the different semantic interoperability processes to the user in an appropriate visualization manner is fundamental. In addition, we should explore new methods to analyse the results of quality of semantic interoperability assessment. Finally, it would be interesting to develop a global semantic interoperability quality measure that integrates the idea of propagation of quality across the different processes that participate in the global semantic interoperability process.

References

Aberer, K., Cudré-Mauroux, P., Hauswirth, M., 2003. The Chatty Web: Emergent Semantics Through Gossiping. In WWW '03: Proceedings of the 12th international conference on World Wide Web, pp.197-206, New York, NY, USA.

Aberer, K., Cudré-Mauroux, P., Ouskel, A.M. *et al.*, 2004. Emergent semantics: principles and issues. In DASFAA 2004, LNCS 2973, Y. Lee *et al.* (Eds.), pp. 25-36 (Berlin: Springer).

Aberer, K., Cudre-Mauroux, Ph. and M. Hauswirth. A framework for semantic gossiping. SIGMOD Record, 31(4), 2002.

Agarwal, P., 2005. Ontological considerations in GIScience. *International Journal of Geographical Information Science,* 19(5), 501-536.

Agostini A., Moro, G., 2004. Identification of Communities of Peers by Trust and Reputation. In Proc. of the 11th Int. AIMSA Conference, Varna, Bulgaria.

Alhashmi, S.M., Siddiqi, J., Revisiting the art of collaboration in the age of internet, *Communication of the IBIMA* (6), P.21-24, 2008.

Allen, J F 1983 Maintaining Knowledge about Temporal Intervals. *Communication of the ACM*, 26(11): 832-843.

Antoniou, G., Franconi, E., and van Harmelen, F., 2005. Introduction to Semantic Web Ontology Languages. In Reasoning Web, Proceedings of the Summer School, Malta, 2005 (Berlin, Heidelberg, New York, Tokyo, 2005), N. Eisinger and J. Maluszynski, (Eds.), LNCS 3564, Springer-Verlag.

D'Aquin, M., Motta, E., Dzbor, M., Gridinoc, L., Heath, T. Sabou, M., 2008. Collaborative Semantic Authoring. *IEEE Intelligent Systems*, pp. 80-83.

Arpinar, I.A., Sheth, A., Ramakrishnan, C., Usery, E.L., Azami, M., Kwan, M.-P., 2006. Geospatial Ontology Development and Semantic Analytics. *Transaction in GIS*, 10(4), 551-575.

Ai, Y., Di, L., Wei, Y., 2009. A taxonomy of geospatial services for global service discovery and interoperability. *Computers & Geosciences*, 2009, doi:10.1016/j.cageo.2007.12.018

Aussenac-Gilles, N., 2005. *Méthodes ascendantes pour l'Ingénierie des connaissances*. Rapport interne, Habilitation à diriger des recherches, Institut de Recherche en Informatique de Toulouse (IRIT), Université Paul Sabatier, Toulouse III.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., 2003. *The Description Logic Handbook*. Cambridge University Press.

Baeten, J.C.M., 2004. A Brief History of Process Algebra. Rapport CSR 04-02, Vakgroep Informatica, Technische Universiteit Eindhoven.

Baeten, J.C.M., Weijland, W.P., 1990. Process Algebra. Number 18 in Cambridge Tracts in Theoretical Computer Science, Cambridge University Press.

Baïna, K., Charoy, F., Godart, C., Grigori, D., El Hadri, S. Skaf, H., 2004. CORVETTE: a cooperative workflow for virtual teams coordination. *Int. J. Networking and Virtual Organisations*, vol. 2, issue 3, pp. 232-245.

Bakillah, M., Mostafavi, M.A., Bédard, Y., 2006. A Semantic Similarity Model for Mapping between Evolving Geospatial Data Cubes. *Proceedings of the Second International Workshop on Semantic-based Geographical Information Systems (SeBGIS06)*, LNCS 4278, Meersman R, Tari Z, Herrero P *et al.* (eds.), Montpellier, France, 1658-1669.

Bakillah, M., Mostafavi, M.A., Brodeur, J., 2011. Multi-View Augmented Concept to Improve Semantic Interoperability of Geospatial Data. In Bo Wu (eds.) ISPRS Book, Taylor & Francis.

Bakillah, M., Mostafavi, M.A., 2010. G-Map Semantic Mapping Approach Based on Augmented Geospatial Service Description to Improve Semantic Interoperability of Distributed Geospatial Web Services. *Fourth International Workshop on Semantic and Conceptual Issues in Geographic Information Systems* (SeCoGIS 2010), J. Trujillo et al. (eds) LNCS 6413, pp. 12-22, November 1-4, Vancouver, Canada.

Bakillah, M., Mostafavi, M.A., Bédard, Y., Brodeur J., 2009a. Elements of Semantic Mapping Quality: a Theoretical Framework. In *Quality Aspects in Spatial Data Mining*. Edited by Alfred Stein, Wenzhong Shi, and Wietske Bijker. CRC Press, 37-45.

Bakillah, M., Mostafavi, M.A., Bédard, Y., Brodeur, J., 2009b. SIM-NET: A View-Based Semantic Similarity Model for Ad Hoc Networks of Geospatial Databases. *Transactions in GIS*, 13(5), 417-447.

Bakillah, M., Mostafavi, M.A., Brodeur, J. Bédard, Y., 2007. Mapping Between Dynamic Ontologies in Support of Geospatial Data Integration for Disaster Management. In Li J Zlatanova S Fabbri A (eds), *Lecture Notes in Geoinformation and Cartography*, Springer Verlag Berlin Heidelberg, 201-224.

Baldauf, M., Schahram, D., Rosenberg, F., 2007. A Survey on Context-Aware Systems. *IJAHUC*, 2, 263-277.

Balram, S., Dragicevic, S., 2006. Modeling Collaborative GIS Processes Using Soft Systems Theory, UML and Object Oriented Design. *Transaction is GIS*, 10(2), 199-218.

Barnes, J., 1954, Class and committees in a Norwegian island parish. *Human Relations* (7), 39-58.

Bédard, Y., Bernier, E., Devillers, R., 2002. La métastructure VUEL et la gestion des représentations multiples, Généralisation et Représentation multiple, Hermes, Vol. Chap. 8, p. 150-162

Bédard, Y., Bernier E., 2002. Supporting Multiple Representations with Spatial View Management and the Concept of "VUEL". *Joint Workshop on Multi-Scale Representations* *of Spatial Data*, ISPRS WG IV/3, ICA Com. on Map Generalization, 7-8 July 2002, Ottawa, Canada.

Benchikha F, Boufaida M, Seinturier L, 2005 Viewpoints: a Framework for Object-Oriented Database Modeling and Distribution. *Data Science Journal* 4: 92-107.

Bennett, B., 2005. Modes of concept definition and varieties of vagueness. *Applied Ontology*, 1(1), pp.17–26.

Ben Messoud, R., Loudcher Rabaséda, S., Missaoui, R., Boussaid, O., 2007. OLEMAR : A Online Environment for Mining Association Rules in Multidimensional Data. *Data Mining and Knowledge Discovery Technologies*, David Taniar (ed.).

Berners-Lee, T., Hendler, J., Lassila, O., 2001. The Semantic Web. *Scientific American*, 34-43.

Bhatt M, Flahive A, Wouters C, Rahayu W, Taniar D, 2006 MOVE: A Distributed Framework for Materialized Ontology Views Extraction. *Algorithmica* 45(3): 457- 481.

Bian, L., Hu, S., 2007. Identifying Components for Interoperable Process Models using Concept Lattice and Semantic Reference System. *International Journal of Geographical Information Science,* 21(9), 1009-1032.

Bishr, Y., 1998. Overcoming the Semantic and other Barriers to GIS Interoperability. *International Journal of Geographical Information Science,* 12(4), 299-314.

Bloehdorn, S., Haase, P., Hefke, M., Sure, Y., Tempich, C., 2005. Intelligent Community Lifecycle Support. In Proc. of the Int. I-KNOW Conference, Graz, Austria.

Boella, G., Van der Torre, L., Villata, S., 2009. Four Measures for the Dynamics of Coalitions in Social Networks. HT'09, Torino, Italy, 361-362.

Bonifacio, M., Bouquet, P., Traverso, P., 2002. Enabling Distributed Knowledge Management. Managerial and Technological Implications. *Novatica and Informatik/Informatique*, 3(1).

Borgida, A., Walsh, T.J., Hirsh, H., 2005. Towards Measuring Similarity in Description Logics. In International Workshop on Description Logics (DL2005). Edinburgh, Scotland.

Bouquet, P., Mikalai, Y., Zanobini, S., 2005. Critical Analysis of Mapping Languages and Mapping Techniques. Technical Report DIT-05-052, University of Trento, Italy.

Bowers, S., Ludäscher, B., "An ontology-driven framework for data transformation in scientific workflow." In LNCS vol. 2994, *Proceedings of the International Workshop on Data Integration in the Life Sciences (DILS'04)*, Leipzig, Germany, 2004.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., 2000. Extensible Markup Language (XML) 1.0 (second edition), W3C Recommendation 6 october 2000.

Brickley, D., Guha, R., 2000. Resource Description Framework Schema Specification 1.0.

Brockmans, S., Haase, P., 2006. A Metamodel and UML Profile for Networked Ontologies: A Complete Reference. Technical report, University Karlsruhe.

Brodaric, B., 2007. Geo-Pragmatics for the Geospatial Semantic Web. *Transactions in GIS,* 11(3), 453-477.

Brodaric, B., Fox, P., McGuiness, D.L., 2009. Geoscience Knowledge Representation in Cyberinfratrusccture. *Computers & Geoscience* 35: 697-699.

Brodeur, J., 2004. Interoperabilité des données géospatiales : élaboration du concept de proximité géosémantique. Ph.D. Thesis, Université Laval, Québec, Canada.

Brodeur, J., Bédard, Y., 2001. Geosemantic Proximity, a Component of geospatial Data Interoperability. In *Internat. Workshop, Semantics of Enterprise Integration, ACM Conference on OOPSLA,* Tampa, Florida: 14–18.

Brodeur, J., Bédard, Y., Edwards, G., Moulin, B., 2003. Revisiting the Concept of Geospatial Data Interoperability within the Scope of Human Communication Process. *Transactions in GIS* 7(2): 243-265.

Brown, J., Broderick, A. J., Lee, N., 2007. Word of Mouth Communication Within Online Communities: Conceptualizing the Online Social Network. *Journal of Interactive Marketing*, 21(3), 2-20.

Carney, D., Smith, J., and Place, P., 2005. Topics in Interoperability: Infrastructure Replacement in a System of Systems (CMU/SEI-2005-TN-031). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University.

Carrington, P.J., Scott, J., Wasserman, S., 2005. *Models and Methods in Social Network Analysis*. Cambridge University Press.

Castano, S., Ferrara, A., Montanelli, S., 2006. Matching Ontologies in Open Networked Systems: Techniques and Applications. *Journal on Data Semantics*, 25-63.

Castano, S., Ferrara, A., Montanelli, S., 2003. H-MATCH: an algorithm for dynamically matching ontologies and peer-based systems. Proceedings of SWDB'03, The first International Workshop on Semantic Web and Databases, pp. 232-251.

Castano, S., Ferrara, A., Montanelli, S., Racca, G., 2004. From Surface to Intensive Matching of Semantic Web Ontologies. In Proc. of the DEXA WEBS 2004 Workshop, Zaragoza, Spain.

Castano, S., Montanelli, S., 2005. Semantic self-formation of communities of peers. Proc. Of the ESWC Workshop on Ontologies in Peer-to-Peer Communities, Heraklion, Greece.

Castillo, J.A.R., Silvescu, A., Caragea, D., Pathak, J., Honavar, V.G., 2003. Information Extraction and Integration from Heterogeneous, Distributed, Autonomous Information Sources – A Federated Ontology-Driven Query-Centric Approach. Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI), pp. 183-191.

Ceglar A., Roddick, J. F., 2006. Association Mining. *ACM Computing Surveys*, 38 (2), 5.

Cilibrasi, R.L., Vitanyi, P.M.B., 2007. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering* 19(3), 370-383.

Congalton, R.G., Green, K., 2009. Assesing the Accuracy of Remotely Sensed Data: Principles and Practices. Second Edition. CRC Press, Taylor and Francis.

Contractor, N. S., Monge, P. R., 2002. Managing knowledge networks. *Management Communication Quarterly*, 16, 249–258.

Contractor, N.S., Wasserman, S., Faust, K., 2006. Testing multitheoretical, multilevel hypotheses about organizational networks: an analytic framework and empirical example. *Academy of Management Review* 31(3), 681-703.

Cormode, G., Muthukrishnan, S., 2007. The String Edit Distance Matching Problem with Moves. ACM Transactions on Algorithms, 3(1).

Couclelis, H., 2003. The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge. *Transactions in GIS*, 7(2), 165-175.

Crespo, A., Garcia-Molina, H., 2002. Semantic Overlay Networks for P2P Systems. Technical report, Computer Science Department, Stanford University.

Cross, V., Pal, A., 2005. Metrics for Ontologies. In Annual Meeting of the NAFIPS, 448-453, Tokyo, Japan.

Cudré-Mauroux, P., 2006. Emergent Semantics: Rethinking Interoperability for Large Scale Decentralized Information Systems. Ph.D. Thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 203 p.

Curé, O., Jeansoulin, R., 2009. An FCA-based Solution for Ontology Mediation. *Journal of Computing Science and Engineering*, 3(2), pp. 90-108.

Curé, O., Jeansoulin, R., 2007. Data Quality Enhancement of Databases Using Ontologies and Inductive Reasoning. In R. Meersman and Z. Tari et al. (Eds.), OTM 2007, Part I, LNCS 4803, pp. 1117–1134.

D'Amato, C., Fanizzi, N., Esposito, F., 2005. A Semantic Similarity Measure for Expressive Description Logics. The Second International Workshop on Knowledge Discovery and Ontologies. Porto, Portugal.

Dang, T.-T., Frankovic, B., Budinská, I., 2003. Optimal Creation of Agent Coalitions for Manufacturing and Control, Computing and Informatics, Slovak Academic Press Ltd., 22(1).

Dey, A.K., 2001. Understanding and Using Context. *Personal and Ubiquitous Computing* 5(1), 4-7.

Devillers, R., Bédard, Y., Jeansoulin, R., Moulin B., 2007. Towards Spatial Data Quality Information Analysis Tools for Experts Assessing the Fitness for Use of Spatial Data. *International Journal of Geographical Information Science,* 21(3), 261-282.

Devillers, R., Bédard, Y., Jeansoulin, R., 2005. Multidimensional Management of Geospatial Data Quality Information for its Dynamic Use within Geographical Information Systems. *American Society for Photogrammetry and Remote Sensing* 71(2), 205-215.

Debrauwer, L. 1998. Des vues aux contextes pour la structuration fonctionnelle de bases de données à objets en CROME. Doctoral thesis, University of sciences and technologies, Lile, France.

Decker, S., Erdmann, M., Fensel, D., Horrocks, I., Klein, M., van Harmelen, F., 2000. OIL in a Nutshell. In Proceedings of EKAW'00, France.

Dholakia, U. M., Bagozzi, R. P., Klein Pearo, L., 2004. A Social Influence Model of Consumer Participation in Network- and Small-group-based Virtual Communities. *International Journal of Research in Marketing*, 21, 241-263.

Ding, H., 2006. *A Semantic Search Framework in Peer-to-Peer Based Digital Libraries*. Ph.D thesis. Norwegian University of Science and Technology.

Ding, Q., Sundarraj, G., 2007. Mining Association Rules from XML Data. *Data Mining and Knowledge Discovery Technologies*, David Taniar (ed.).

Do, H.H., Melnik, S., Rahm, E., 2003. Comparison of Schema Matching Evaluation. In LNCS 2593, ed. A.B. Chaudhri *et al.*, 221-237. Berlin, Heidelberg: Springer-Verlag.

Do, H.H., Rahm, E., 2001. COMA- A System for Flexible Combination of Schema Matching Approaches. In Proc. of the 28th Conf. on VLDB, 610-621. Hong Kong, China.

Do, H.-H., Rahm, E., 2002. COMA: a System for Flexible Combination of Schema Matching Approaches. In Proceedings of the 28th International Conference on Very Larges Databases. Hong Kong, China: 610-621.

Doan, A.-H., Madhavan, J., Domingos, P., Halevy, A., 2004. Ontology Matching: A Machine Learning Approach. In Steffen Staab and Rudi Studer (eds.), *Handbook on Ontologies*, Springer Verlag, 385–516.

Duchon, P., Hanusse, N., Lebhar, E., Schabanel, N., 2006. Towards Small World Emergence. Proceedings of the eighteenth annual ACM symposium on Parallelism in Algorithms and Architectures (SPAA '06). ACM. New York, NY, USA, pp. 225-232.

Ebers, M., 1999, The dynamics of inter-organizational relationships. *Research in the sociology of organizations*, 16, 31-56.

Egenhofer, M.J. 2002. Toward the Semantic Geospatial Web. *GIS'02*, November 8-9, 2002, McLean, Virginia, USA.

Egenhofer M 1993 A Model for Detailed Binary Topological Relationships. *Geomatica*, 47(3 & 4): 261-273.

Ehrig, M., Staab, S., 2004. QOM – Quick Ontology Matching. The Semantic Web ISWC 2004.

Euzenat J. et al. State of the art on current alignment techniques. KnowledgeWeb Deliverable 2.2.3, 2004. http://knowledgeweb.semanticweb.org.

Euzenat, J., Shvaiko, P., 2007. *Ontology Matching*. Springer-Verlag: Berlin Heidelberg.

Euzenat, J., Valtchev, P. 2004. Similarity-Based Ontology Alignment in OWL-lite. In Proceedings 15th European Conference on Artificial Intelligence (ECAI), 333-337.

Fallahi, G.R., Frank, A.U., Mesgari, M.S., and Rajabifard, A., 2008. An Ontological Structure for Semantic Interoperability of GIS and Environmental Modeling. *International Journal of Applied Earth Observation and Geoinformation*. 10(3): 342-357.

Farquhar, A., Fikes, R., Rice, J., 1996. The Ontolingua Server: A Tool for Collaborative Ontology Construction. In Proceeding of the 10th Knowledge Acquisition for Knowledge-based Systems workshop.

Farrugia, J.A., 2007. Semantic Interoperability of Geospatial Ontologies: A Model theoretic Analysis. Ph.D. Thesis, University of Maine.

Fensel, D., Crubezy, M., van Harmelen, F., Horrocks, I., 2000. Oil & upml: A Unifying Framework for the Knowledge Web. In proceedings of ECAI 2000, Berlin, Germany.

Fensel, D., Hendler, J., Lieberman, H., Wahlster, W., 2003. *Spinning the Semantic Web*. MIT Press.

Fernandez, J.I.R., 2007. Semantic Interoperability in Ad Hoc Computing Environments. Ph.D. thesis, Loughborough University.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, et al., 2001. Placing search in context: the concept revisited. Proceedings of the 10th international conference on World Wide Web, pp. 406-414.

Firat, A., Madnick, S., Grosof, B., 2007. Contextual Alignment of Ontologies in the Ecoin Semantic Interoperability Framework. *Information Technology and Management*, 8(1), 47-63.

Fonseca, F., Camara, G., and Monteiro, A.M., 2005. A Framework for Measuring the Interoperability of Geo-Ontologies. *Spatial Cognition and Computation,* 6(4), 307-329.

Fox, S., 2008. Ontological Uncertainty and Semantic Uncertainty in Global Network Organizations. VTT Working Papers.

Fu, G. H., Jones, C. B., Abdelmoty, A. I., 2005. Ontology-based spatial query expansion in information retrieval. In Meersman, R. and Tari, Z. (eds.). Proceedings of On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE, Part 2. Berlin, Springer Lecture Notes in Computer Science No 3762, pp. 1466–82.

Fulk, J., Flanigin, A. J., Kalman, M. E., Momge, P. R., Ryan, T. 1996. Connective and communal public goods in interactive communication systems. *Communication Theory*, 6, 60–87.

Ganter B, Wille R, 1999 Formal *Concept Analysis: Mathematical Foundations*. Springer, Berlin.

Gesbert, N., 2005. *Study of the Formalization of Geographical Database Specifications for their Integration*. Ph.D. Thesis. Université de Marne-la-Vallée, France.

Giuchiglia, F. Shvaiko, P., 2004. Semantic Matching. *Knowledge Engineering Review*, 18(3), 265–280.

Giuchiglia, F. Shvaiko, P., Yatskevich, M., 2004. S-Match: An Algorithm and an Implementation of Semantic Matching. Proceedings of the European Semantic Web Symposium, LNCS 3053, 61-75.

Giunchiglia, F., Zaihrayeu, I., 2002. Making Peer Databases Interact - A Vision for an Architecture Supporting Data Coordination. 6th International Workshop on Cooperative Information Agents (CIA), Madrid, Spain.

Goodchild, M.F., 2007. Beyond Metadata: Towards User-centric Description of Data Quality. Paper read at Proceedings, Spatial Data Quality 2007 (International Symposium on Spatial Data Quality, June 13-15, at Enschede, Netherlands.

Goodchild, M.F., Egenhofer, M.J. and Fegeas, R., 1998. Interoperating GISs: Report of the Specialist Meeting, Santa Barbara, CA, National Center for Geographic Information and Analysis, pp. 1–27.

Grenon, P. and Smith, B., 2004. SNAP and SPAN: Towards Dynamic Spatial Ontology. *Journal of Spatial Cognition and Computation*, 14(1), 69 – 104.

Gruber, T.R., 1993. A Translation Approach to Portable Ontology Specification. Stanford, California, Knowledge Systems Laboratory Technical Report KSL 92-71.

Hafsia, R., 2001. *Semantic interoperability in ad hoc wireless networks*. Ph. D. thesis. Naval Postgraduate School, Monterey, California.

Haase, P., Siebes, R., van Harmelen, F., 2008. Expertise-based Peer Selection in Peer-to-Peer Networks. *Knowledge and Information Systems*. London: Springer, 75-107.

Harvey F, Kuhn W, Pundt H, Bishr Y, Riedemann C, 1999 Semantic interoperability: a central issue for sharing geographic information. *Annals of Regional Science* (33): 213-232.

Heath, T., 2008, Information Seeking on the Web with Trusted Social Networks – from Theory to Systems. Ph. D. thesis. Knowledge Media Institute, The Open University.

Hess, G.N., Iochpe, C. Castano, S., 2007. Geographic Ontology Matching with IG Match. In SSTD 2007, LNCS 4605, D. Papadias, D. Zhang, and G. Kollios (Eds.), 185–202.

Hoare, C.A.R., 1978. Communicating Sequential Processes. *Communications of the ACM*, 21(8), 666–677.

Hochmair H 2005 Ontology matching for spatial data retrieval from internet portals. In Rodríguez M A, Cruz I F, Egenhofer M J, and Levashkin S (eds) *Proceedings of the First International Geospatial Semantics Conference.* Berlin, Springer Lecture Notes in Computer Science No 3799: 166–82

Horrocks, I., Patel-Schneider, P., Boley, H., Tabet, S., Grosof, B., Dean, M., 2004. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. Available at http://www.w3.org/Submission/SWRL.

Horrocks, I., Patel-Schneider, P. F., van Harmelen, F., 2002. Reviewing the Design of DAML+OIL: An Ontology Language for the Semantic Web. In Proceedings of AAAI 2002, pp. 792–797.

Hoser, B., Hotho, A., Jäschke, R., Schmitz, C., Stumme, G., Semantic Network Analysis of Ontologies. *KDML 2006*, p. 297-305, 2006.

Hossein, M., Rajabifard, A., Williamson, I.P., 2010. Development of an Interoperable Tool to Facilitate Spatial Data Integration in the Context of SDI. *International Journal of Geographical Information Science*, 24(4), 487 – 505.

Hu, W., Zhao, Y., Li, D., Cheng, G., Wu, H., Qu, Y., 2007. Falcon-AO: Results for OAEI 2007. Proceedings of the International Workshop on Ontology Matching.

Hu, W., Qu, Y., 2008. Falcon-AO: A Practical Ontology Matching System. *Web Semantics: Science, Services and Agents on the World Wide Web* 6, 237–239.

Janowicz, K., 2006. Sim-DL: Towards a semantic similarity measurement theory for the description logic ALCNR in geographic information retrieval. In *Proceedings of the Second International Workshop on Semantic-based Geographical Information Systems (SeBGIS06)*, LNCS 4278, Meersman R, Tari Z, Herrero P *et al.* (eds), Montpellier, France: 1681-1692.

Janowicz, K., Raubal, M., Schwering, A., Kuhn, W., 2008. Semantic Similarity Measurement and Geospatial Application — Editorial Guest. *Transaction in GIS*.

Jian, W.H.N., Cheng, G., Qu, Y., 2005. Falcon-AO: Aligning Ontologies with Falcon. In *Proceedings K CAP Workshop on Integrating Ontologies*, 87–93.

Jung, J.J., Juszczyszyn, K., 2007. Centrality measurement on semantically multiplex social network: divide-and-conquer approach. *International Journal of Information and Database Systems*, 1(3-4), 277-292.

Kalfoglou, Y., Schorlemmer, M., 2003. Ontology Mapping: the State of the Art. *The Knowledge Engineering Review,* 18 (1), 1–31.

Kantere, V., Tsoumakos, D., Sellis, T., A framework for semantic grouping in P2P databases, *Information Systems* 33, p. 611-636, 2008.

Kavouras, M., A unified ontological framework for semantic integration, in: *Proc. Int. Workshop on Next generation Geospatial Information*, 2003, 19–21.

Kavouras, M., Kokla, M., 2008. *Theories of geographic concepts*. CRC Press, Taylor & Francis Group.

Kavouras, M., Kokla M., and Tomai E., 2005. Comparing Categories among Geographic Ontologies. *Computers & Geosciences*, 31,145–154.

Keeney, J., Lewis, D., O'Sullivan, D., Roelens, A., Wade, V., Boran, A., Richardson, R., 2006. Runtime Semantic Interoperability for Gathering Ontology-based Network Context. *Network Operations and Management Symposium, 2006*, 10th IEEE/IFIP, 56-65.

Kent, S., 1997. Constraint diagrams: Visualizing invariants in object oriented modelling. In *Proceedings of OOPSLA97*, pages 327-341. ACM Press.

Keßler, C., Raubal, M., Janowicz, K., 2007. The Effect of Context on Semantic Similarity Measurement. *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*, 1274-1284.

Khambatti, M., Ryu, K.D., Dasgupta, P., 2002. Peer-to-peer Communities: Formation and Discovery. *14th IASTED Conference on Parallel and Distributed Computing Systems (PDCS)*, Cambridge, Massachusetts, 166-173.

Kifer, M., Lausen, G., Wu., J., 1995. Logical Foundations of Object-Oriented and Frame-based Languages. *Journal of the ACM.*

Klien, E., Lutz, M., and Kuhn, W., 2006. Ontology-Based Discovery of Geographic Information Services - An Application in Disaster Management. *Computers, Environment and Urban Systems* (CEUS), 30, 102-123.

Koh, Y.S., O'Keefe, R., Rountree, N., 2007. Current Interestingness Measures for Association Rules: What Do They Really Measure? *Data Mining and Knowledge Discovery Technologies*, David Taniar (ed.).

Kopetz, H., 2011. Real-Time Systems: Design Principles for Distributed Embedded Applications. Real-Time System Series. Springer Science+Business Media.

Krackhardt, D., 1994. Constraints on the interactive organization as an ideal type. In C. Heckscher & A. Donnellon (Eds.), *The post-bureaucratic organization: New perspectives on organizational change*, 211–222. ThousandOaks, CA: Sage.

Kuhn, W., 2003. Semantic Reference System. *International Journal of Geographical Information Science,* 17(5), 405-409.

Kuhn, W., 2005. Geospatial Semantics: Why, of What, and How? *Journal on Data Semantics*, *Special Issue on Semantic-based Geographical Information Systems*, 3534: 1-24.

Kuhn, W., Raubal, M., 2003. Implementing Semantic Reference Systems. AGILE 2003 - 6th AGILE Conference on Geographic Information Science, Collection des sciences appliquées de l' INSA de Lyon, M. Gould, R. Laurini, and S. Coulondre, Eds. Lyon, France: Presses Polytechniques et Universitaires Romandes, pp.63-72.

Lambert, M., 2006. Développement d'une approche pour l'analyse SOLAP en temps réel : adaptation aux besoins des activités sportives en plein air. Master Thesis, Université Laval, Québec, Canada.

Lawrence, S., 2000. Context in Web Search. *IEEE Data Engineering Bulletin* 23: 25-32.

Lee, Y-W., Park, H.-H., Shibasaki, R., 2006. Collaborative GIS Environment for Exploratory Spatial Data Analysis Based on Hybrid P2P Network. Z. Pan et al. (Eds.) Edutainment 2006, LNCS 3942, 330-333.

Lemmens, R. 2006. Semantic Interoperability of Distributed Geo-Services. Ph.D Thesis, International Institute for Geo-Information Science and Earth Observation (ITC), Enschede, The Netherlands, 323 p.

Lemmens, R., Wytzisk, A., de By, R., Granell, C., Gould, M. and P. van Osterom, 2006. Integrating Semantic and Syntactic Descriptions to Chain Geographic Services. *IEEE Internet Computing*, 10(5), 42-52.

Li, Y., Zuhair, A., McLean, D., 2003. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering* 15(4): 871–882.

Liu, K., Bhaduri, K., Das, K., Nguyen, P., Kargupta, H., 2006. Clientside web mining for community formation in peer-to-peer environments. WEBKDD'06, Philadelphia, Pennsylvania, USA.

Lopez, V., Sabou, M., and Motta, E., 2006. PowerMap: Mapping the Real Semantic Web on the Fly. In ISWC 2006, LNCS 4273, I. Cruz et al. (Eds.), pp. 414 – 427.

Löser, A., Naumann, F., Siberski, W., Nejdl, W., Thaden, U., 2003. Semantic Overlay Clusters Within Super-Peer Networks. *DBISP2P*, 33-47.

Lumineau, N., Doucet, A., 2004. Sharing Communities Experiences for Query Propagation in Peer-to-Peer Systems. Proceedings of the International Database Engineering and Application Symposium (IDEAS'04), 1-8.

Lutz, M., 2005. Ontology-Based Service Discovery in Spatial Data Infrastructures. Proceedings of the 2nd Workshop on Geographic Information Retrieval, Bremen, Germany, 45-54.

Lutz, M., Klien, E., 2006. Ontology-based retrieval of geographic information. *International Journal of Geographical Information Science* 20: 233–60.

Lutz, M., Riedemann, C., Probst, F., 2003. A Classification Framework for Approaches to Achieving Semantic Interoperability Between GI Web Services. In *Spatial Information Theory: Foundations of Geographic Information Science*, LNCS 2825, Kuhn, W., Worboys, M.F., and Timpf, S., (eds) Berlin: Springer : 186-203.

Madhavan, J., Bernstein, P., Rahm, E., 2001. Generic Schema Matching with Cupid. In Proc. of the 28th Conf. on VLDB, 49-58. Hong Kong, China.

Maedche, A., Staab, S., 2002. Measuring Similarity between Ontologies. In Proc. of Int. Conf. on Knowledge Engineering and Knowledge Management, 251-263, Siguenza, Spain.

Maguitman, A., Menczer, F., Roinestad, H., Vespignani, A., 2005. Algorithmic Detection of Semantic Similarity. International WWW Conference, Chiba, Japan: 107-116.

Mandreoli, F., Martoglia, R., Penzo, W., Sassatelli, S., 2006. SRI: Exploiting semantic information for effective query routing in a PDMS. Proceedings of the 8th ACM International Workshop on Web Information and Data Management (WIDM 2006), Arlington, Virginia, USA.

Manso, M. M., Wachowicz, M., 2009. GIS Design: A Review of Current Issues in Interoperability. *Geography Compass*, 3(3), 1105-1124.

Mao, M. Peng, Y., 2006. PRIOR System: Results for OAEI 2006. Proceedings of the International Workshop on Ontology Matching (OM-2006) collocated with the 5th

International Semantic Web Conference (ISWC-2006), volume 225 of CEUR Workshop Proceedings, 173–180.

Margolis, E., Laurence, S. (Eds.), 1999. *Concepts Core Readings*. The MIT Press.

Massmann, S., Engmann, D., Rahm, E., 2006. COMA++: Results for the Ontology Alignment Contest OAEI 2006. Proceedings of the International Workshop on Ontology Matching (OM-2006) collocated with the 5th International Semantic Web Conference (ISWC-2006), volume 225 of CEUR Workshop Proceedings, 107–114.

McGuinness, D., van Harmelen F. (eds.), 2003. OWL Web Ontology Language Overview W3C Proposed Recommendation.

Medin, D.L., Rips, L.J., 2005. Concepts and Categories: Memory, meaning and metaphysics, concepts and categorization. In *The Cambridge Handbook of Thinking and Reasoning*. Ed. K.J. Holyoak and R.G. Morrisson, pp. 37-72. Cambridge, UK: Cambridge University Press.

Mika, P., 2006. Ontologies Are Us: a Unified Model of Social Networks and Semantics. *Journal of Web Semantics* 5, 5-15.

Milner, R., 1980. A Calculus of Communicating Systems. Lecture Notes in Computer Science No. 92, Springer Verlag.

Melnik, S., Garcia-Molina, H., Rahm, E., 2002. Similarity Flooding: A Versatile Graph Matching Algorithm and its Application to Schema Matching. Proceedings of the 18th International Conference on Data Engineering (ICDE02), 117–128.

Miller G 1995 WordNet: A Lexical Database for English. Communications of the ACM, 38(11): 39–41.

Mostafavi, M. A., 2006. Semantic Similarity Assessment in Support of Geospatial Data Integration. In 7th Int. Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 685-693. Lisbon, Portugal.

Mostafavi, M.A., Edwards, G., Jeansoulin, R., 2004. An Ontology-Based Method for Quality Assessment of Spatial Data Bases. In ISSDQ'04 Proc., 49-66. Bruck am der Leitha, Austria.

Mostowfi, F., Fotouhi, F., 2006. Improving Quality of Ontology: An Ontology Transformation Approach. In Proc. of the 22nd ICDEW, 61, Miami, USA.

Mika, P., 2004. Social Networks and the Semantic Web. In Proc. of the IEEE/WIC/ACM Int. WI Conference, Beijing, China.

Mika, P., 2006. Ontologies Are Us: a Unified Model of Social Networks and Semantics. *Journal of Web Semantics* 5, 5-15.

Montanelli, S., Castano, S., 2008. Semantically Routing Queries in Peer-based System: the H-LINK Approach. *The Knowledge Engineering Review*, 23:51-72. Cambridge University Press.

Nagy, M., Vargas-Vera, M., Motta, E., 2006. DSSim-ontology Mapping with Uncertainty. Proceedings of the International Workshop on Ontology Matching (OM-2006) collocated with the 5th International Semantic Web Conference (ISWC 2006), volume 225 of CEUR Workshop Proceedings, 115–123.

Nejdl, W., Wolpers, M., Siberski, W., Schmitz, C., Schlosser, M.T., Brunkhorst, I., Löser, A., 2004. Super-peer-based Routing Strategies for Rdf-based Peer-to-peer Networks. *J. Web Sem.*, 1(2), 177-186.

Niedbala, S. OWL-CtxMatch in the OAEI 2006 alignment contest. In Proceedings of the International Workshop on Ontology Matching (OM-2006) collocated with the 5th International Semantic Web Conference (ISWC-2006), volume 225 of CEUR Workshop Proceedings, pages 165–172, November 2006.

Nittel, S., Duckham, M., Kulik, L., 2004. Information Dissemination in Mobile Ad-Hoc Geosensor Networks. Geographic Information Science, LNCS 3234, Springer: Berlin, 206-222.

Noy, N.F., Musen, M.A., 2001. Anchor-Prompt: Using Non-Local Context for Semantic Matching. In Proc. of Workshop on Ontologies and Information Sharing at International Joint Conference on Artificial Intelligence, 63-70. Seattle, USA.

Noy, N.F., Musen, M.A., 2004. Specifying Ontology Views by Traversal. In Proceedings of the Third International Conference on the Semantic Web (ISWC-2004), LNCS 3298: 713-725.

Obrst, L., 2003, Ontologies for Semantically Interoperable Systems. In Proceedings of the 12th international conference on information and knowledge management, New Orleans, LA, USA, 366-369.

Ogden, C.K. and I.A. Richards, *The Meaning of Meaning*. 1946: Harcourt, Brace & World.

Omran, E.E., van Etten, J., 2007. Spatial-Data Sharing: Applying Social-Network Analysis to Study Individual and Collective Behaviour. *International Journal of Geographical Information Science*, 21(6), 699-714.

Osman, N., Robertson, D., 2007. Dynamic Verification of Trust in Distributed Open Systems. IJCAI, 1440-1445.

Ovarec, V., Budinska, I., Frankovic, B., 2007. Coalition Representation in Ontology Using Various Types of Logics. 5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics, Poprad, Slovakia, 49-58.

Papamarkos, G., Poulovassilis, A., Wood, P.T., 2003. Event-Condition-Action Rule Languages for the Semantic Web.

Parent, C., Spaccapietra, S., Zimaniyi, E., 2006. The MurMur Project: Modeling and Querying Multi-Representation Spatiotemporal Databases. *Information Systems*, 31(8), 733-769.

Park, J., Ram, S., 2004. Information Systems Interoperability: What Lies Beneath? *ACM Transactions on Information Systems,* 22(4), 595-632.

Pathak, N., Mane, S., Srivastana, J., 2007. Analysis of Cognitive Social and Knowledge Networks from Electronic Communication. *International Journal of Semantic Computing,* 1(1), 87-120.

Pavard, B., 2002. An Introduction to Complexity in Social Science. Toulouse: Université P. Sabatier.

Preece, J., Maloney-Krichmar, D., 2005. Online Communities: Design, Theory, and Practice. *Journal of Computer-Mediated Communication*, 10(4), 00.

Poole, M.S., 1999. Organizational Challenges for the New Forms. In G. DeSanctis & J. Fulk (Eds.), *Shaping organizational form: Communication, connection, and community*, 453–471. Thousand Oaks, CA: Sage.

Rada, R., Mili, H., Bicknell, E., Blettner, M., 1989. Developement and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics* 19(1), 17-30.

Raftopoulou, P., Petrakis, E., 2005. Semantic similarity measures: a comparison study. Technical University of Crete, Intelligent Systems Laboratory, Technical Report TR-TUC-ISL-04 2005.

Raubal, M. 2004. Formalizing Conceptual Spaces. In Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS 2004), Torino, Italy, 153-164.

Rawat, S. 2003. Interoperable Geo-Spatial Data Model in the Context of the Indian NSDI. Enschede, The Netherlands: International Institute for Geo-information Science and Earth Observation.

Resnick, P. 1999. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research* 11, 95-130.

Robertson, D., 2004. A Lightweight Coordination Calculus for Agents Systems. Second International Workshop on Declarative Agent Languages and Technologies (DALT), Columbia University, New York.

Rodriguez, A., Egenhofer, M., 2003, Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering,* 15 (2), 442–456.

Sadiq, M. Z., Duckam, M., 2009. Integrated storage and querying of spatially varying data quality information in a relational spatial database. *Transactions in GIS*, 13(1), 25-42.

Sauro, L., 2005. Formalizing Admissibility Criteria in Coalition Formation Among Goal Directed Agents. PhD thesis, University of Turin.

Schwering, A., 2008. Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS,* 12(1), 5-29.

Schwering, A., Raubal, M., 2005. Measuring Semantic Similarity Between Geospatial Conceptual Regions. In Proceedings of the First International Conference on GeoSpatial Semantics, LNCS 3799, Mexico City, Mexico, 90-106.

Schwering, A., Kuhn, W., 2009. An Hybrid Semantic Similarity Measure for Spatial Information Retrieval. *Spatial Cognition & Computation: An Interdisciplinary Journal* 9(1), 30 – 63.

Serafini, L., Bouquet, P., Magnini, B., Zanobini, S., 2003. An Algorithm for Matching Contextualized Schemas Via SAT. Technical Report # 0301−06, Istituto Trentino di Cultura, Trento, Italy.

Serrat, O., 2009. Social Network Analysis. *Knowledge Solutions* No. 28.

Shvaiko, P., 2004. A Classification of Schema-Based Matching Approaches. In Proceedings of the Meaning Coordination and Negotiation Workshop at the ISWC'04, 119-130.

Shadbolt, N., Hal, W., and Berners-Lee, T., 2006. The Semantic Web revisited. *IEEE Intelligent Systems* 21, 96–101.

Smirnov, A., Pashkin, M., Chilov, N., Levashova T., 2006. Context-Based Disaster Management Support. Knowledge Systems for Coalition Operation (KSCO).

Smith, B., Mark, D., 1998. Ontology and Geographic Kinds. International Symposium on Spatial Data Handling, Vancouver, Canada, 308-320.

Staab, S., Tempich, C., and Wranik, A. 2004. REMINDIN': Semantic query routing in peer-to-peer networks based on social metaphors. In Proceedings of the 13th International conference on World Wide Web (WWW 2004), New York, NY, USA.

Staub, P., Gnägi, H.R., Morf, A., 2008, Semantic Interoperability Through the Definition of Conceptual Model Transformation. *Transactions in GIS* 12(2), 193-207.

Steffens T 2005 Knowledge-Rich Similarity-Based Classification. *Proceedings of ICCBR 2005*, LNCS 3620, Munoz-Avila H and Ricci F(eds), Chicago, USA, Springer: 522-536

Stuckenschmidt, H., 2003. Ontology-Based Information Sharing in Weakly Structured Environments. PhD thesis, Vrije Universiteit, Amsterdam.

Stuckenschmidt H 2006 Toward Multi-viewpoint Reasoning with OWL Ontologies. Y. Sure and J. Domingue (Eds.): ESWC 2006, LNCS 4011, pp. 259–272.

Sun Microsystems, 2010. JXTA v2.3.x: JavaTM Programmer's Guide. JXTA project documentation, www.jxta.org.

Tan, P.S., Goh, A.E.S., Lee, S.S.G., A context model for B2B collaborations, *IEEE International Conference on Service Computing*, p. 108-15, 2008.

Tang, J., Li, J., Liang, B., Huang, X., Li, Y., Wang, K., 2006. Using Bayesian Decision for Ontology Mapping. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(4), 243–262.

Tomai, E., Kavouras, M., 2004. From “Onto-GoeNoesis” to “Onto-Genesis”: The Design of Geographic Ontologies. *GeoInformatica*, 8(3), pp.285-302.

Tryfona, N., Price R., Jensen C.S., 2003, Conceptual Models for Spatiotemporal Applications. I In: T. Sellis *et al*, ed. *Spatiotemporal Databases: the Chorochronos Approach.* LNCS *2520, Berlin, Springer, pp. 79-116.*

Tversky, A., 1977. Features of Similarity. *Psychological Review* 84(4): 327-352.

Vaccari, L., Shvaiko, P., Marchese, M., 2009. A Geo-service Semantic Integration in Spatial Data Infrastructures. *International Journal of Spatial Data Infrastructures Research*, 4, 24-51.

van der Torre, L., Villata, S., 2009. Four Ways to Change Coalitions: Agents, Dependencies, Norms and Internal Dynamics. SNAMAS ’09, 19–24.

Wand, Y., Wang, R., 1996. Anchoring Data Quality Dimensions in Ontological Foundation. *Communications of the ACM* 39(11): 86-95.

Wasserman, S., Faust, K., 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Wiegand N., Garcia C. 2007 A Task-Based Ontology Approach to Automate Geospatial Data Retrieval. Transaction in GIS 11(3), 355-376.

Worboys, M., Duckham, M., 2006. Monitoring Qualitative Spatiotemporal Change for Geosensor Networks. International Journal of Geographical Information Science, 20(10), 1087 – 1108.

Wouters, C., Dillon, T.S., Rahayu, W., Meersman, R., Chang, E., 2008. Extraction Process Specification for Materialized Ontology Views. T.S. Dillon et al. (Eds.): Advances in Web Semantics I, LNCS 4891: 130–175.

Xiao, L., Hu, B., Hederman, L., Lewis, P., Dimitrov, B.D., Fahey, T., 2009. Towards Knowledge Sharing and Patient Privacy in a Clinical Decision Support System. Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, 99-104.

Yang, B, Garcia-Molina, H., 2002. Efficient Search in Peer-to-Peer Networks. In ICDCS 2002.

Yu, C.Q., Peuquet, D., 2008. A GeoAgent-based Framework for Knowledge-oriented Representations: Embracing Social Rules in GIS. *International Journal of Geographical Information Science,* iFirst Article, 1–38.

Zafeiropoulos, A., Spanos, D.-E., Arkoulis, S., Konstantinou, N., Mitour, N. 2009. Data Management in Sensor Networks Using Semantic Web Technologies. In H. Jin and Z. Lv, Data Management in Semantic Web. Nova Science Publishers.

Zaihrayeu, I., 2006. *Towards peer-to-peer information management systems*. PhD thesis, International Doctorate School in Information and Communication Technology, University of Trento, Italy.

Zambonelli, F., Jennings, N.R., Wooldrige, M., 2003. Developing Multi-agents Systems: the Gaia Methodology. *ACM Transactions on Software Engineering and Methodology*, 12, 317–370.

Zeinalipour-Yazti, D., Kalogeraki, V., Gunopulos, D., 2005 Exploiting Locality for Scalable Information Retrieval in Peer-to-Peer Networks. *Information Systems*, 30(4), 277–298.

Zhang, D., Li, Y., Li, J., Tang, J., 2006. Result of Ontology Alignment with RiMOM at OAEI‘06. In Proceedings of the International Workshop on Ontology Matching (OM-2006) collocated with the 5th International Semantic Web Conference (ISWC 2006), volume 225 of CEUR Workshop Proceedings, 181–190.

Zhao, H., 2007. Semantic Matching Across Heterogeneous Data Sources. *Communication of the ACM*, 50(1), 45–50.

Zhdanova, A., 2008. Community-driven ontology construction in social networking portals. *Web Intelligence and Agent Systems: An International Journal*, vol. 6, pp.1-29.

Zheng, L., Tang, J., Jin, Z., 2008. Requirement Driven Service Agent Coalition Formation and Negotiation. The 9th International Conference for Young Computer Scientists, 322-329.

Zhuge, H., Liu, J., Feng, L., He, C., 2004. Semantic-based Query Routing and Heterogeneous Data Integration in Peer-to-Peer Semantic Link Networks. Proceedings of the International Conference on Semantics of a Networked World (ICSNW), in Cooperation with ACM SIGMOD 2004, 91–107, Paris, France.

Annex 1: Excerpt of OWL Contexts of Geospatial Databases Used in the Experimentation

```
<geospatialDatabase rdf:ID="Administrative_and_Topographic_Database">
        <has_domain rdf:resource="#airport"/>
        <has_domain rdf:resource="#designated_area"/>
        <has_domain rdf:resource="#hydrography"/>
        <has_domain rdf:resource="#road_network"/>
        <has_geographical_location rdf:resource="#quebec"/>
        <has_function rdf:resource="#identify_road_status"/>
        <has_function rdf:resource="#identify_topological_relations"/>
        <has_function rdf:resource="#localise_airport"/>
        <has_function rdf:resource="#localise_dam"/>
        <has_function rdf:resource="#localise_island"/>
        <has_function rdf:resource="#localise_urban_area"/>
        <has_function rdf:resource="#localise_waterbody"/>
        <has_function rdf:resource="#localise_watercourse"/>
        <has_time_period_end rdf:datatype="&xsd;dateTime"
            >2011-10-22T00:00:00</has_time_period_end>
        <has_time_period_start rdf:datatype="&xsd;dateTime"
            >2011-01-01T00:00:00</has_time_period_start>
    </geospatialDatabase>
    <geospatialDatabase rdf:ID="Canadian_Disaster_Database">
        <has_domain rdf:resource="#biological_disaster"/>
        <has_domain rdf:resource="#geological_disaster"/>
        <has_domain rdf:resource="#human_disaster"/>
        <has_domain rdf:resource="#hydrological_disaster"/>
        <has_domain rdf:resource="#meteorological_disaster"/>
        <has_domain rdf:resource="#technological_disaster"/>
        <has_geographical_location rdf:resource="#canada"/>
        <has_function rdf:resource="#identify_disaster"/>
        <has_function rdf:resource="#identify_material_dammage"/>
        <has_function rdf:resource="#identify_number_of_victims"/>
        <has_function rdf:resource="#localise_disaster"/>
        <has_function rdf:resource="#localise_material_dammage"/>
        <has_function rdf:resource="#maintain_history_of_disaster"/>
        <has_time_period_end rdf:datatype="&xsd;dateTime"
            >2011-10-22T00:00:00</has_time_period_end>
        <has_time_period_start rdf:datatype="&xsd;dateTime"
            >1900-01-01T00:00:00</has_time_period_start>
    </geospatialDatabase>
```

Annex 2: Excerpt of OWL Contexts of Coalitions Produced by the Coalition Discovering Component

```
    <coalition rdf:ID="coalition_for_evacuation_planning">
        <has_domain rdf:resource="#road_network"/>
        <has_geographical_location rdf:resource="#canada"/>
        <has_member rdf:resource="#Demographic_Database_of_Canada"/>
        <has_member
rdf:resource="#National_Search_and_Rescue_Secretariat_Database"/>
        <has_member rdf:resource="#National_Topographic_Database"/>
        <has_function rdf:resource="#assess_population_density"/>
        <has_function rdf:resource="#evacuation_planning"/>
        <has_function rdf:resource="#identify_road_classification"/>
        <has_time_period_end rdf:datatype="&xsd;dateTime"
            >2011-10-22T00:00:00</has_time_period_end>
        <has_time_period_start rdf:datatype="&xsd;dateTime"
            >1950-01-01T00:00:00</has_time_period_start>
    </coalition>
    <coalition rdf:ID="coalition_for_monitoring_toxic_waste">
        <has_domain rdf:resource="#industry"/>
        <has_geographical_location rdf:resource="#canada"/>
        <has_member rdf:resource="#National_Topographic_Database"/>
        <has_member rdf:resource="#Toxic_Waste_Monitoring_Database"/>
        <has_function
rdf:resource="#produce_chemicals_accident_risk_map"/>
        <has_time_period_end rdf:datatype="&xsd;dateTime"
            >2011-10-22T00:00:00</has_time_period_end>
        <has_time_period_start rdf:datatype="&xsd;dateTime"
            >2010-01-01T00:00:00</has_time_period_start>
    </coalition>
    <coalition rdf:ID="coalition_to_assess_flooding_risk">
        <has_domain rdf:resource="#hydrography"/>
        <has_domain rdf:resource="#hydrological_disaster"/>
        <has_geographical_location rdf:resource="#canada"/>
        <has_member rdf:resource="#Canadian_Disaster_Database"/>
        <has_member
rdf:resource="#Environment_Canada_Meteorological_Database"/>
        <has_member rdf:resource="#National_Topographic_Database"/>
        <has_member rdf:resource="#Natural_Risk_Database"/>
        <has_function rdf:resource="#assess_flooding_risk"/>
        <has_function rdf:resource="#localise_flood"/>
        <has_time_period_end rdf:datatype="&xsd;dateTime"
            >2011-10-22T00:00:00</has_time_period_end>
        <has_time_period_start rdf:datatype="&xsd;dateTime"
            >1900-01-01T00:00:00</has_time_period_start>
    </coalition>
```

Annex 3: Excerpt of Ontologies Developed for the Implementation

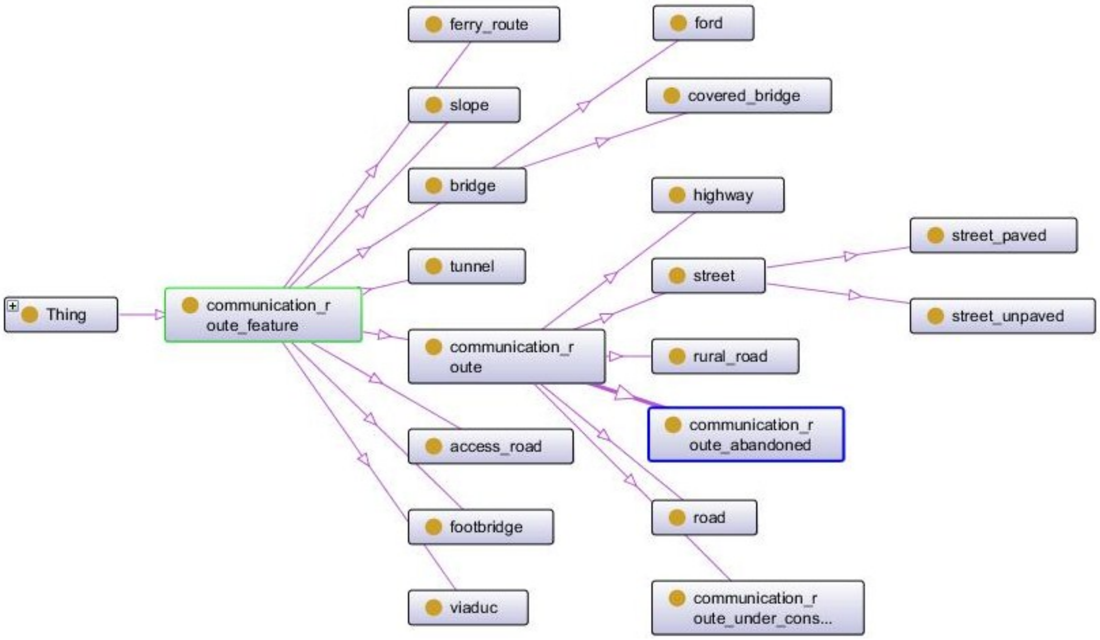


Figure A3.1 BDTQ hydrographic features

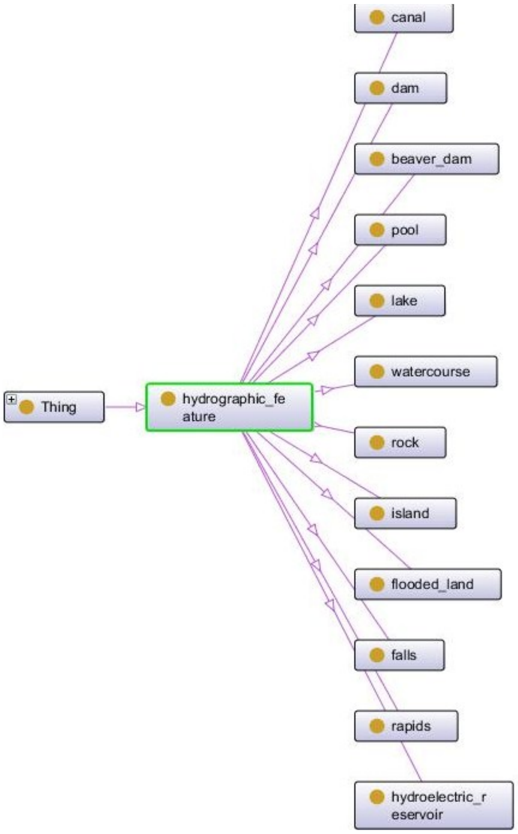


Figure A3.2 BDTQ hydrographic features

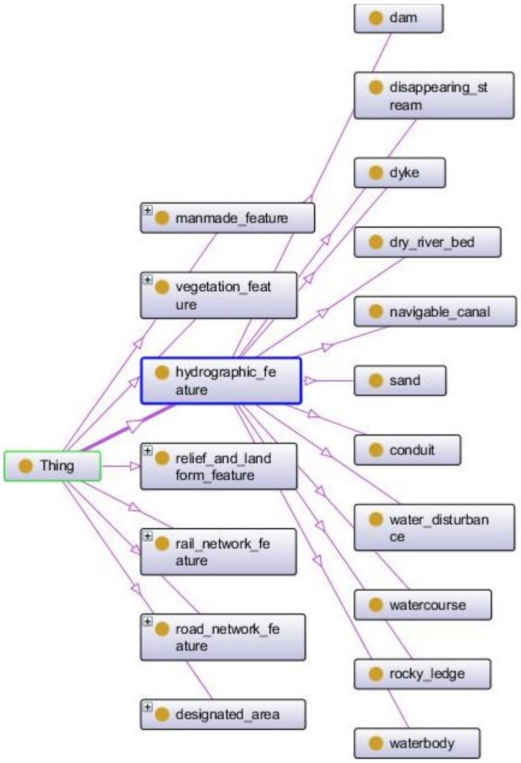


Figure A3.3 NTDB hydrographic features

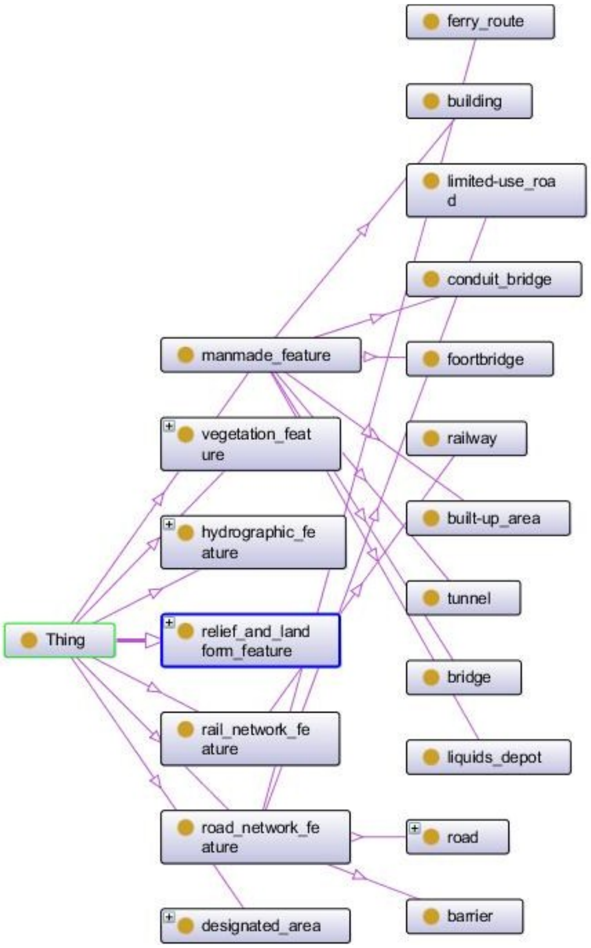


Figure A3.4 NTDB road network

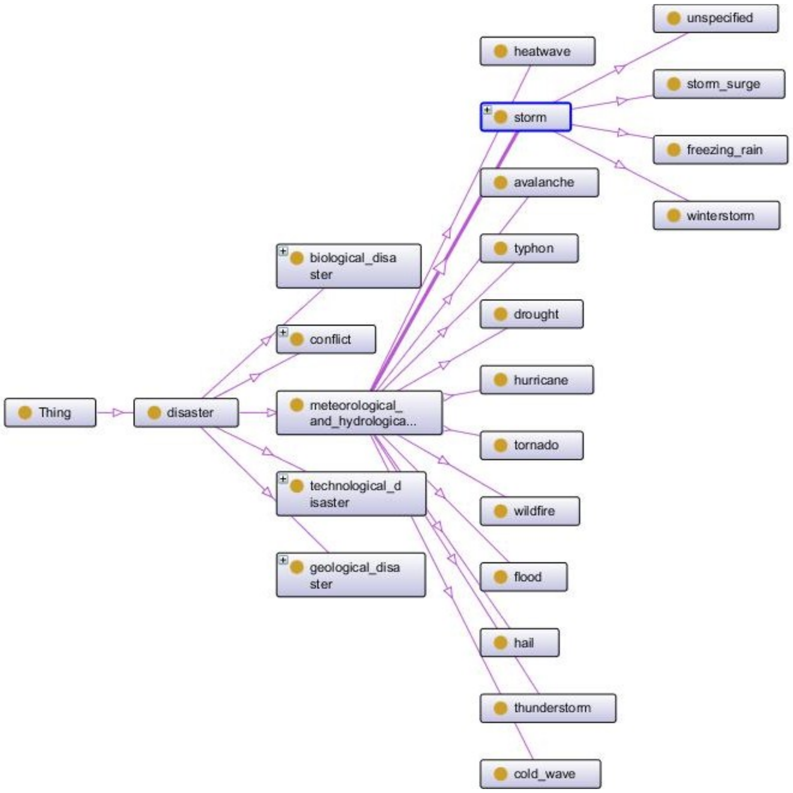


Figure A3.6 Types of disasters in the Canadian Disaster Database (part 1, meteorological and hydrological disaster)

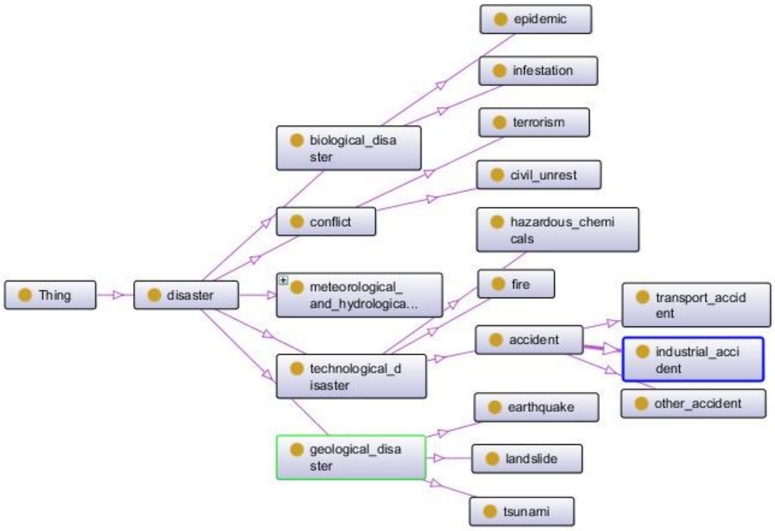


Figure A3.7 Types of disasters in the Canadian Disaster Database (part 2)

Annex 4: Example of Augmented Concept “Road” from BDTQ

```
<MVACconcept>
            <idconcept>101</idconcept>
            <nomconcept>road</nomconcept>
            <augmentedview>
                  <contexte>functional context: for the displacement of
vehicles</contexte>
                  <proprietes>
                <enspropprietesthem>
                  <proprietethem>
                        <nom>surface</nom>
                        <ensval>
                            <val>paved</val>
                        </ensval>
                    </proprietethem>
                    <proprietethem>
                        <nom>status</nom>
                        <ensval>
                            <val>praticable</val>
                        </ensval>
                    </proprietethem>
                    <proprietethem>
                        <nom>road classification</nom>
                        <ensval>
                            <val>highway</val>
                            <val>street</val>
                            <val>rural road</val>
                        </ensval>
                    </proprietethem>
                </enspropprietesthem>
                <enspropprietesspat>
                    <proprietespat>
                        <nom>geometry</nom>
                        <ensval>
                            <val>line</val>
                            <val>polygon</val>
                        </ensval>
                    </proprietespat>
                </enspropprietesspat>
                <enspropprietestemp>
                </enspropprietestemp>
            </proprietes>
            <relations>
                <ensrelthem>
                  <relthem>
                        <nom>is-a</nom>
                        <memb1>road</memb1>
                        <memb2>road network feature</memb2>
                    </relthem>
                </ensrelthem>
                <ensrelspat>
```

```
      <relspat>
            <nom>overlap</nom>
             <memb1>road</memb1>
             <memb2>bridge</memb2>
        </relspat>
        <relspat>
            <nom>in</nom>
             <memb1>road</memb1>
             <memb2>rural area</memb2>
             <memb2>urban area</memb2>
        </relspat>
    </ensrelspat>
    <ensreltemp>
    </ensreltemp>
</relations>
<descripteurs>
      <ensdescrspat>
            <descrspat>
                  <nom>geometry represents</nom>
            <ensval>
                <val>central axis of roadway</val>
                <val>roadway</val>
            </ensval>
            </descrspat>
            <descrspat>
                  <nom>width</nom>
            <ensval>
                <val>less than 20m</val>
                <val>20m and over</val>
            </ensval>
            </descrspat>
            <descrspat>
                  <nom>shape</nom>
            <ensval>
                <val>curve</val>
            </ensval>
            </descrspat>
      </ensdescrspat>
      <ensdescrtemp>
      </ensdescrtemp>
</descripteurs>
      <dependencies>
            <dep>
                  <head>
                        <concept1>road</concept1>
                        <char1>surface</char1>
                        <ensval1>
                              <val>paved</val>
                        </ensval1>
                  </head>
                  <body>
                        <concept2>road</concept2>
                        <char2>status</char2>
                        <ensval2>
```

```
                                         <val>praticable</val>
                                   </ensval2>
                             </body>
                             </dep>
                        <dep>
                             <head>
                                   <concept1>road</concept1>
                                   <char1>surface</char1>
                                   <ensval1>
                                         <val>unpaved</val>
                                   </ensval1>
                             </head>
                             <body>
                                   <concept2>road</concept2>
                                   <char2>status</char2>
                                   <ensval2>
                                         <val>not praticable</val>
                                   </ensval2>
                             </body>
                             </dep>
                  </dependencies>
            </augmentedview>
            <augmentedview>
                  <contexte>situational context: in
construction</contexte>
                  <proprietes>
                tensproprietesthem>
                  <proprietethem>
                        <nom>surface</nom>
                        <ensval>
                            <val>unpaved</val>
                        </ensval>
                    </proprietethem>
                    <proprietethem>
                        <nom>status</nom>
                        <ensval>
                            <val>not praticable</val>
                        </ensval>
                    </proprietethem>
                    <proprietethem>
                        <nom>road classification</nom>
                        <ensval>
                            <val>highway</val>
                            <val>street</val>
                            <val>rural road</val>
                        </ensval>
                    </proprietethem>
                </ensproprietesthem>
                <ensproprietesspat>
                    <proprietespat>
                        <nom>geometry</nom>
                        <ensval>
                            <val>line</val>
                            <val>polygon</val>
            </ensval>
        </proprietespat>
    </ensproprietesspat>
    <ensproprietestemp>
    </ensproprietestemp>
</proprietes>
<relations>
    <ensrelthem>
      <relthem>
            <nom>is-a</nom>
            <memb1>road</memb1>
            <memb2>road network feature</memb2>
        </relthem>
    </ensrelthem>
    <ensrelspat>
      <relspat>
            <nom>overlap</nom>
             <memb1>road</memb1>
             <memb2>bridge</memb2>
        </relspat>
        <relspat>
            <nom>in</nom>
             <memb1>road</memb1>
             <memb2>rural area</memb2>
             <memb2>urban area</memb2>
        </relspat>
    </ensrelspat>
    <ensreltemp>
    </ensreltemp>
</relations>
<descripteurs>
      <ensdescrspat>
            <descrspat>
                  <nom>geometry represents</nom>
            <ensval>
                <val>central axis of roadway</val>
                <val>roadway</val>
            </ensval>
            </descrspat>
            <descrspat>
                  <nom>width</nom>
            <ensval>
                <val>less than 20m</val>
                <val>20m and over</val>
            </ensval>
            </descrspat>
            <descrspat>
                  <nom>shape</nom>
            <ensval>
                <val>curve</val>
            </ensval>
            </descrspat>
      </ensdescrspat>
      <ensdescrtemp>
      </ensdescrtemp>
</descripteurs>
      <dependencies>
            <dep>
                  <head>
                        <concept1>road</concept1>
                        <char1>surface</char1>
                        <ensval1>
                              <val>unpaved</val>
                        </ensval1>
                  </head>
                  <body>
                        <concept2>road</concept2>
                        <char2>status</char2>
                        <ensval2>
                              <val>not praticable</val>
                        </ensval2>
                  </body>
                  </dep>
      </dependencies>
</augmentedview>
<augmentedview>
      <contexte>situational context: abandoned</contexte>
      <proprietes>
    <ensproprietesthem>
      <proprietethem>
            <nom>surface</nom>
            <ensval>
                <val>paved</val>
                <val>unpaved</val>
            </ensval>
        </proprietethem>
        <proprietethem>
            <nom>status</nom>
            <ensval>
                <val>not praticable</val>
            </ensval>
        </proprietethem>
        <proprietethem>
            <nom>road classification</nom>
            <ensval>
                <val>highway</val>
                <val>street</val>
                <val>rural road</val>
            </ensval>
        </proprietethem>
    </ensproprietesthem>
    <ensproprietesspat>
        <proprietespat>
            <nom>geometry</nom>
            <ensval>
                <val>line</val>
                <val>polygon</val>
            </ensval>
        </proprietespat>
    </ensproprietesspat>
    <ensproprietestemp>
    </ensproprietestemp>
</proprietes>
<relations>
    <ensrelthem>
      <relthem>
            <nom>is-a</nom>
            <memb1>road</memb1>
            <memb2>road network feature</memb2>
        </relthem>
    </ensrelthem>
    <ensrelspat>
      <relspat>
            <nom>overlap</nom>
             <memb1>road</memb1>
             <memb2>vegetation</memb2>
             <memb2>bridge</memb2>
        </relspat>
        <relspat>
            <nom>in</nom>
             <memb1>road</memb1>
             <memb2>rural area</memb2>
             <memb2>urban area</memb2>
        </relspat>
    </ensrelspat>
    <ensreltemp>
    </ensreltemp>
</relations>
<descripteurs>
      <ensdescrspat>
            <descrspat>
                  <nom>geometry represents</nom>
            <ensval>
                <val>central axis of roadway</val>
                <val>roadway</val>
            </ensval>
            </descrspat>
            <descrspat>
                  <nom>width</nom>
            <ensval>
                <val>less than 20m</val>
                <val>20m and over</val>
            </ensval>
            </descrspat>
            <descrspat>
                  <nom>shape</nom>
            <ensval>
                <val>curve</val>
            </ensval>
            </descrspat>
      </ensdescrspat>
      <ensdescrtemp>
```

```
            </ensdescrtemp>
      </descripteurs>
            <dependencies>
                  <dep>
                        <head>
                              <concept1>road</concept1>
                              <char1>overlap</char1>
                              <ensval1>
                                    <val>vegetation</val>
                              </ensval1>
                        </head>
                        <body>
                              <concept2>road</concept2>
                              <char2>status</char2>
                              <ensval2>
                                    <val>not praticable</val>
                              </ensval2>
                        </body>
                        </dep>
                  <dep>
                        <head>
                              <concept1>road</concept1>
                              <char1>surface</char1>
                              <ensval1>
                                    <val>unpaved</val>
                              </ensval1>
                        </head>
                        <body>
                              <concept2>road</concept2>
                              <char2>status</char2>
                              <ensval2>
                                    <val>not praticable</val>
                              </ensval2>
                        </body>
                        </dep>
            </dependencies>
      </augmentedview>

</MVACconcept>
```