LUND UNIVERSITY

Modeling and improving Spatial Data Infrastructure (SDI)

Abdolmajidi, Ehsan

2016

Modeling and improving Spatial Data Infrastructure (SDI)

EHSAN ABDOLMAJIDI

DEPARTMENT OF PHYSICAL GEOGRAPHY AND ECOSYSTEM SCIENCES | LUND UNIVERSITY

Legend
- - - VGI
— NVDB

0 0.15 0.3   0.6   0.9   1.2
Kilometers

Legend
Urban area
Rural area
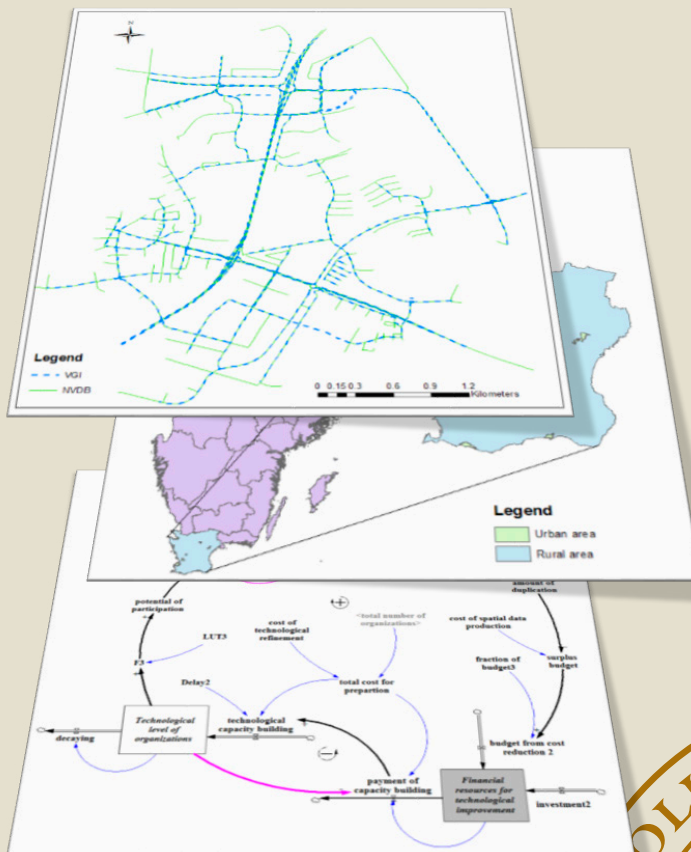
Modeling and improving
Spatial Data Infrastructure (SDI)

Modeling and improving
Spatial Data Infrastructure (SDI)

Ehsan Abdolmajidi

Organization LUND UNIVERSITY Department of Physical Geography and Ecosystem Science, GIS center, Sölvegatan 12, SE-22362 LUND	Document name DOCTORAL DISSERTATION
	Date of issue
Author(s) Ehsan Abdolmajidi	Sponsoring organization The European Union funding program Erasmus Mundus Action-2 (lot 9)

Title and subtitle
Modeling and improving Spatial Data Infrastructure (SDI)

Abstract:

Spatial Data Infrastructure (SDI) development is widely known to be a challenging process owing to its complex and dynamic nature. Although great effort has been made to conceptually explain the complexity and dynamics of SDIs, few studies thus far have actually modeled these complexities. In fact, better modeling of SDI complexities will lead to more reliable plans for its development. A state-of-the-art simulation model of SDI development, hereafter referred to as SMSDI, was created by using the system dynamics (SD) technique. The SMSDI enables policy-makers to test various investment scenarios in different aspects of SDI and helps them to determine the optimum policy for further development of an SDI. This thesis begins with adaption of the SMSDI to a new case study in Tanzania by using the community of participant concept, and further development of the model is performed by using fuzzy logic. It is argued that the techniques and models proposed in this part of the study enable SDI planning to be conducted in a more reliable manner, which facilitates receiving the support of stakeholders for the development of SDI.

Developing a collaborative platform such as SDI would highlight the differences among stakeholders including the heterogeneous data they produce and share. This makes the reuse of spatial data difficult mainly because the shared data need to be integrated with other datasets and used in applications that differ from those originally produced for. The integration of authoritative data and Volunteered Geographic Information (VGI), which has a lower level structure and production standards, is a new, challenging area. The second part of this study focuses on proposing techniques to improve the matching and integration of spatial datasets. It is shown that the proposed solutions, which are based on pattern recognition and ontology, can considerably improve the integration of spatial data in SDIs and enable the reuse or multipurpose usage of available data resources.

Key word: Spatial Data Infrastructure, System Dynamics, Fuzzy Logic, Data integration, Pattern detection, Resource Description Framework (RDF), Ontology
Classification system and/or index terms (if any)

Supplementary bibliographical information	Language: English
Avhandlingar från Institutionen för naturgeografi och ekosystemvetenskap	

ISSN and key title	ISBN (print): 978-91-85793-65-5 ISBN (PDF): 978-91-85793-66-2

Recipient's notes	Number of pages	Price
	Security classification	

Signature

Date 2016-10-11

Modeling and improving
Spatial Data Infrastructure (SDI)

Ehsan Abdolmajidi

LUND
UNIVERSITY

A doctoral thesis at the university in Sweden is produced either as monograph or as a collection of papers. In the latter case the introduction part constitutes the formal thesis, which summarizes the accompanying papers already published or manuscripts at various stages (in press, submitted, or in preparation)

To my beloved father and mother

Content

Abstract

Spatial Data Infrastructure (SDI) development is widely known to be a challenging process owing to its complex and dynamic nature. Although great effort has been made to conceptually explain the complexity and dynamics of SDIs, few studies thus far have actually modeled these complexities. In fact, better modeling of SDI complexities will lead to more reliable plans for its development. A state-of-the-art simulation model of SDI development, hereafter referred to as SMSDI, was created by using the system dynamics (SD) technique. The SMSDI enables policy-makers to test various investment scenarios in different aspects of SDI and helps them to determine the optimum policy for further development of an SDI. This thesis begins with adaption of the SMSDI to a new case study in Tanzania by using the community of participant concept, and further development of the model is performed by using fuzzy logic. It is argued that the techniques and models proposed in this part of the study enable SDI planning to be conducted in a more reliable manner, which facilitates receiving the support of stakeholders for the development of SDI.

Developing a collaborative platform such as SDI would highlight the differences among stakeholders including the heterogeneous data they produce and share. This makes the reuse of spatial data difficult mainly because the shared data need to be integrated with other datasets and used in applications that differ from those originally produced for. The integration of authoritative data and Volunteered Geographic Information (VGI), which has a lower level structure and production standards, is a new, challenging area. The second part of this study focuses on proposing techniques to improve the matching and integration of spatial datasets. It is shown that the proposed solutions, which are based on pattern recognition and ontology, can considerably improve the integration of spatial data in SDIs and enable the reuse or multipurpose usage of available data resources.

1 Introduction

Designing a reliable strategic plan for developing a Spatial Data Infrastructure (SDI) is difficult because the SDI has a complex and dynamic nature in which the components influence each other over time (Chan and Williamson 1999; Erik de Man 2006; Hendriks, Dessers, and van Hootegem 2012). SDI is an infrastructure used to facilitate efficient and effective spatial data management, access, sharing, usage, and reusage among a network of stakeholders (Hendriks, Dessers, and van Hootegem 2012; Vandenbroucke et al. 2009; Hjelmager et al. 2008). In such a collaborative and heterogeneous data sharing environment, spatial data integration is a challenging problem as well. Hence, this thesis is being developed upon a two-fold topic: *Modeling and Improving SDI*.

The first part of this PhD thesis focuses on modeling the complexity and dynamics of SDI development in order to provide valuable insights for policy-makers. Significant effort has been made to conceptually explain the complexity and dynamics of SDIs (Chan et al. 2001; Erik de Man 2006; Grus, Crompvoets, and Bregt 2006; Grus, Crompvoets, and Bregt 2010), whereas few studies have actually modeled the SDI's complexities. In fact, better modeling of the SDI complexities enables more reliable plans to be drawn for its development. This study begins by employing a state-of-the-art simulation model of SDI development, hereafter referred to as SMSDI, which was proposed by Mansourian and Abdolmajidi (2011). The SMSDI is one of the few models that actually represents the complexity and dynamics of an SDI over time. This study integrates the SMSDI with the community of participant concept in order to adapt the model to the new case study in Tanzania. The SMSDI is then further improved by employing fuzzy logic in order to better model the vagueness and uncertainties of the qualitative factors in the SMSDI.

In the second part of this thesis, novel methodologies are proposed to improve the integration of authoritative spatial datasets and Volunteered Geographic Information (VGI) in the SDI environment. Different types of data heterogeneities impede the integration of datasets from different resources, particularly when two datasets are integrated and matched at the instance level. The issue is further highlighted when VGI sources are involved in the data integration because VGIs are often not well-structured at that level. This part of the research begins by proposing a matching algorithm that was enhanced by a pattern detection method.

Although the matching results were significantly improved, it is revealed that pattern detection methods and geometric measures are insufficient for handling complex situations. Afterward, a Resource Description Framework (RDF) data model is proposed that can store explicit description of the features and the relationships among them. This information is key in determining the correct matching pairs in the proposed methodology. To collect data in such a data structure, a simple supervisory mechanism is also suggested for employment in VGI projects.

1.1 Aims

This PhD thesis has two overall aims. The first aim is to employ the state-of-the-art model for SDI development and then to improve the simulation model to better represent the linguistic variables in two case studies. The second aim is to propose a methodology to improve the state of data integration in an SDI environment. The detailed objectives of this thesis are as follows:

- Suggesting an effective approach to better involve national organizations in the SDI planning process (Paper I).

- Improving the representation of the qualitative factors in the SDI development simulation model (Paper II).

- Providing an efficient and effective algorithm to match road network datasets (Paper III).

- Improving the structuring of the road data to facilitate the data integration (Paper IV).

- Providing a simple strategy for collecting well-structured road network data in a VGI project (Paper V).

1.2 Thesis Structure

The thesis is organized in five chapters. The second chapter describes the SDI concept, system dynamics (SD) technique, and fuzzy logic. The related background is also thoroughly presented in this chapter. Chapter 3 elaborates on the geospatial data integration and the available approaches in road network matching. In addition, the results of previous studies are explored. Chapter 4 summarizes the papers produced in this PhD research. Finally, the overall conclusions from this PhD thesis are presented in chapter 5.

The following papers are presented in this thesis:

Paper I   Mansourian, A., Lubida, A., Pilesjö, P., **Abdolmajidi, E.** and Lassi, M., 2015. SDI planning using the system dynamics technique within a community of practice: lessons learnt from Tanzania. *Geo-spatial Information Science*, *18*(2-3), pp. 97-110.

Paper II   **Abdolmajidi, E.**, Harrie, L. and Mansourian, A., 2016. The stock-flow model of spatial data infrastructure development refined by fuzzy logic. *SpringerPlus*, *5*(1), pp. 1-20.

Paper III   **Abdolmajidi, E.**, Mansourian, A., Will, J. and Harrie, L., 2015. Matching authority and VGI road networks using an extended node-based matching algorithm. *Geo-spatial Information Science*, *18*(2-3), pp. 65-80.

Paper IV   **Abdolmajidi, E.**, Harrie, L. and Mansourian, A., 2016. Well-structured Road Data in RDF to Facilitate the Instance Matching: Matching VGI and Authoritative Road Data (Manuscript).

Paper V   **Abdolmajidi, E.**, Harrie, L. and Mansourian, A., 2016. Investigating the Feasibility of Collecting Well-Structured VGI (Manuscript).

In paper I, EA developed the simulation models and helped in processing the results. EA also helped in preparing the questionnaire.

In paper II, EA conducted the data collection, developed the model, and processed the results. He is the main author of the paper.

In paper III, EA prepared the data, developed the algorithm, and assessed the results. He is the main author of the paper.

In paper IV, EA developed the data model and the matching algorithm, and he analyzed the results. He is the main author of the paper.

In paper V, EA developed the Web application and analyzed the results. He is the main author of the paper.

Paper I and III are reproduced with kind permission from the copyright holder.

2 SDI

The SDI concept has emerged as a result of the increasing numbers of multi-participant environments in decision making, which highlights the need to reorganize data across different disciplines and organizations. An SDI is built upon a geospatial data community with a hierarchical structure; hence, it inherits such properties, as shown in Figure 1.
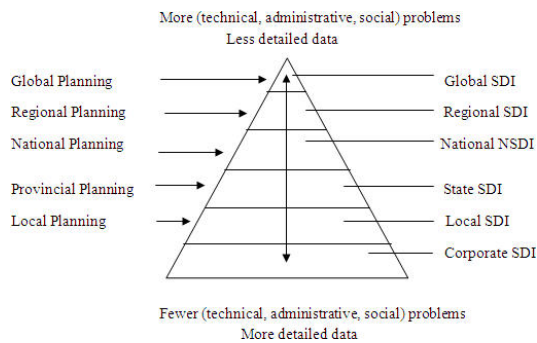
Figure 1. SDI hierarchical structure (Rajabifard 2001).

To implement an SDI, which includes various interacting components (Figure 2) (Rajabifard, Feeney, and Williamson 2002), a variety of institutional, technological, economic, and political factors are involved (Crompvoets et al. 2004; Groot and McLaughlin 2000). These factors have feedback and timely interactions that render an SDI as a complex adaptive system (Grus, Crompvoets, and Bregt 2010) requiring a long-term implementation plan.
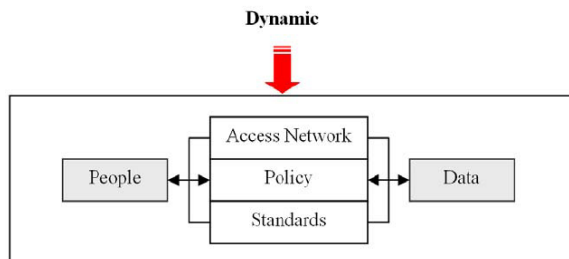
Figure 2. SDI core components (Rajabifard, Feeney, and Williamson 2002).

2.1 SDI modeling

Various models have been developed and proposed in order to simplify the SDI complexity and to gain insight into the nature and behavior of SDIs (Rajabifard and Williamson 2003; Omran, Crompvoets, and Bregt 2006; Grus, Crompvoets, and Bregt 2006; De Man 2007). All of these models provide the rigid foundation necessary for understanding the concept and nature of SDIs. Such foundations enable the coordinating agencies to better plan and implement SDIs, and help researchers to develop more intuitive models.

A framework based on a network perspective was suggested by Vandenbroucke et al. (2009) to characterize an SDI. In this model, the main players in a spatial data community are identified, and the data flow between them is explained. The authors used Social Network Analysis (SNA) to assess the applicability of their framework by measuring "density," "distance," and "centrality" parameters of the network of data producers and users. The results showed that the proposed framework is suitable for representing the dataflow among stakeholders. It is also useful in analyzing the behavior of different types of role players in an SDI.

The Commission on Spatial Data Standards of the International Cartographic Association (ICA) has also defined and proposed a framework according to the ISO Reference Model for Open Distributed Processing (RM-ODP) standard. This framework is composed of a set of formal conceptual models, each of which targets one characteristic of SDIs (Hjelmager et al. 2005; Cooper et al. 2007; Hjelmager et al. 2008; Cooper et al. 2013). This framework describes complex distributed systems in different levels of abstraction (Delgado 2004), such as Enterprise, Information, Computation, Engineering, and Technology (Hjelmager et al. 2005). The purpose, scope, and policies of an SDI are described in the Enterprise view. The Information view models the semantics and information processes associated with the SDI. In the Computation model, the SDI is then decomposed into objects and services interacting at the interface level. The Engineering perspective contains mechanism and functions necessary to incorporate the distributed interactions among objects in an SDI. The last viewpoint, Technology, describes the specific technologies required for implementing the SDI.

Hjelmager et al. (2005) developed initial models of the first two viewpoints by using Unified Modeling Language (UML). They described the scope, activities, and actors from the Enterprise perspective and the information semantics and processes of the Information viewpoint. Hjelmager et al. (2005) distinguished the following five different roles of stakeholders: 1) producer, 2) provider, 3) broker, 4) user, and 5) policy-maker. In their later article, the value-added reseller role was added (Hjelmager et al. 2008). The producer only produces the data or service in

an SDI; while the provider provides the data and service to the user. The broker is the specialized publishers who maintain the metadata and help users and providers to negotiate on contracts between them. The value-added reseller produces new products by adding new features to the existing products and offers them to the users. The last is the policy-maker, who sets policies mandating other role players to pursue them. Data and services are considered as products in the Enterprise and Information models. In the Information viewpoint, the policy is the initial point, which determines the specification for the products. Other elements of the Information model are metadata and catalogue along with information and knowledge derived from the data. Hjelmager et al. (2008) then connected each stakeholder to each of these elements. The stakeholders can be passively or actively involved in any element of the Information viewpoint model.

Cooper et al. (2007) also developed an initial model for the Computational perspective that captures the details of the services and interfaces regardless of their distribution, which is or should be covered in the Engineering viewpoint. The ICA commission proposed that an SDI has six different types of services: Registry, Data, Processing, Portrayal, Application, and Management. They used the Component diagram of UML to represent the objects and services in an SDI. In this model, the required and implemented interfaces are also represented as relationships among these services. Such a model is an appropriate means for verifying whether the implemented system has the required components. A more comprehensive SDI model from the Computational viewpoint was then presented by Cooper et al. (2013). In their article, the services and interfaces are connected to the stakeholders identified in the Enterprise viewpoint. These models provide more holistic representation of an SDI independent from any legislation, technology, and implementation.

Mansourian and Abdolmajidi (2011) also developed a state-of-the-art simulation model using the SD technique. Their model is able to represent the complex dynamic characteristics of the SDI while engaging all participating components in the SDI development. They argued that SMSDI can provide policy-makers with deeper insight into the outcomes of their plans for developing an SDI. They detected four growth engines associated with the main components of an SDI (Figure 2). A growth engine is a positive feedback loop that can improve the state of the system with little initial value (investment). The growth engine for the data-sharing culture represents SDI awareness of high-level managers and low-level operators. The managers are involved in policy-making and prioritizing their involvement in the SDI development; low-level operators are responsible for data production and storage and are directly involved in sharing data in the SDI framework. The standards growth engine then depicts the process of improving the standards while different organizations join the SDI development. The technological level of organizations is also reflected in the technology growth

engine. The technological level of organizations includes the overall status of all organizations in terms of having proper equipment for producing, storing, and sharing data (through network access). Mansourian and Abdolmajidi (2011) tested the model and used it for different scenarios involving different policies of investment in SDI. The simulation results showed promising accordance to the actual situation in their case study, and the tests results demonstrated the reliability of SMSDI.

2.2 Community of practice

The community of practice (Lave and Wenger 1991) concept was first used to describe an apprenticeship whereby novices learn a profession from experts through observation and practice (Lave and Wenger 1991). The community is defined as inter- or intra-organizational groups of people who are often geographically dispersed and have been working on knowledge-sharing or knowledge-creating activities. Each of these groups with distinct identities focuses on a certain practice such as a professional discipline, skill, or topic (Verburg and Andriessen 2006). The community of practice describes collaborative learning of community members of a professional or organizational setting. Strengths and weaknesses of individual community members in different areas highlight the collaborative learning process as a vital aspect of this model. A community of practice has a common domain of interest that defines its identity and has a shared practice introducing communication techniques so that the members of community can collaboratively interact and learn from each other (Wenger, McDermott, and Snyder 2002). Because the community of practice brings people from different fields together, a common ground should be defined in order for the community to succeed (Clark and Brennan 1991). That is, the community members need to share mutual beliefs, knowledge, and common vocabulary (Wenger 2000). Establishing a common ground is essential for sharing the community's knowledge with other communities and organizations and also for "developing a shared understanding of complex systems of ideas that the community develops (Ahmad and Al-Sayed 2005)."

Paper I begins with the development of a simulation model for an SDI using the SD technique. The SDI development, in this study, is considered as a community of practice in order to promote the interactive learning among different involved organizations and policy-makers. In the context of the SDI development, experts from stakeholder organizations across the case study country or region build the groups in the community of practice framework. They, as a formal expert community (Verburg and Andriessen 2006), are gathered to develop a common

understanding of SDI, which finally leads to an SDI development model agreed upon by all.

2.3 The system dynamics technique

Designing better policies for developing complex systems requires tools and processes that can help to understand such complexities (Sterman 2000). SD is a technique for modeling systems composed of complex, dynamic, and nonlinear feedback networks. This method gives insight into the collective behavior of a complex system by using simulation models regarding different policies for developing or changing the system. Such insight reduces the development cost and increases the reliability of system development because the advantages and disadvantages of the policies can be detected before implementation.

The simulation model is built using the stock-flow model, which allows the modeler to integrate qualitative and quantitative variables in the model and to calculate their feedback (Forrester 1958, 1961, 1968, 1969; Sterman 2000). A stock-flow simulation model is composed of stocks, flows, and auxiliary variables, as described below (Figure 3):
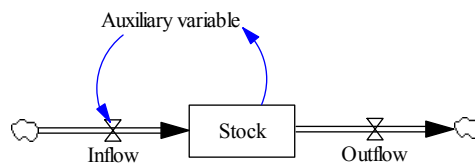
Figure 3. A simple stock-flow structure.

- Stocks represent the states of a system every moment and are timely regulated by in- and out-flow variables.

- Flows are variables controlling the flow into and out of the stocks based on the discrepancy between the desired and current states of affairs and institute corrective action.

- Auxiliary variables and constants are complementary elements that calculate or provide required information in the system.

The SD techniques has been used to model numerous complex systems such as urban, industrial, and ecological systems (Dudley and Soderquist 1999; Forrester 1961, 1969) as well as information technology (Dai, Xiao, and Xie; Quaddus and Intrapairot 2008). This technique is able to bring qualitative and quantitative variables of a system together in order to model the dynamic behavior of the

system in a feedback system over time. However, the qualitative variables create uncertainties when they are quantified in the model.

2.4 Fuzzy logic

Owing to the exclusive characteristics of a humanistic system such as an SDI, accurate quantitative analyses of its behavior are likely irrelevant because conventional quantitative techniques of system analysis are incompatible for addressing humanistic systems (Zadeh 1973). Zadeh (1973) suggested an alternative approach based on fuzzy set labels, which are the key elements in human thought processes, rather than numbers. A human is able to summarize information into linguistic characteristics (labels) that are relevant to the task at hand. For instance, a manager may perceive the financial status of his or her organization as "strong" or "weak" (two labels) to qualitatively evaluate the organization, which may be compared with that of other organizations. Therefore, fuzzy logic can better reflect human judgment in SD models by providing fuzzy inputs and decision rules. This has motivated researchers in the field of system thinking to apply the concept of fuzzy SD in their simulation models. Accordingly, many components in an SDI are directly or indirectly influenced by human decision based on the deployed development plan, which creates uncertainties in SMSDI. Therefore, fuzzy logic can improve the representation of such humanistic characteristics of an SDI by better modeling the linguistic variables in the SMSDI.

Three main steps are used to implement fuzzy logic in a simulation model (Figure 4): fuzzification, inference, and defuzzification.
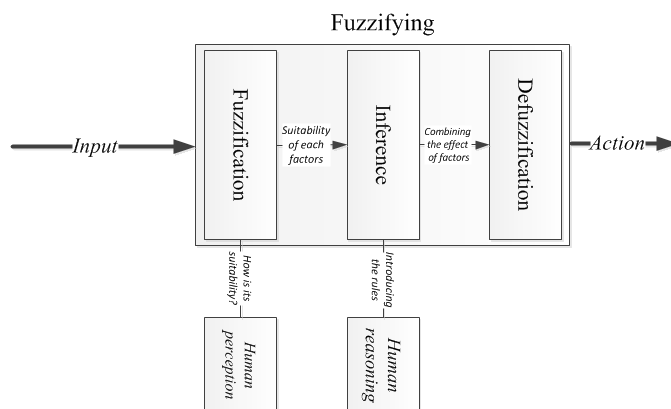
Figure4. Fuzzifying steps.

The first step is to calculate the level of belonging of the current state to each label (Labib, Williams, and O'Connor 1998) by using predefined fuzzy membership functions. Figure 5 shows three fuzzy membership functions for the labels regarding the cultural level of organizations: low, medium, and high. An organization involved in SDI development has a membership value to each of these labels at a time stamp $t0$. The membership value can then change over time owing to different policies such as holding workshops for increasing awareness.
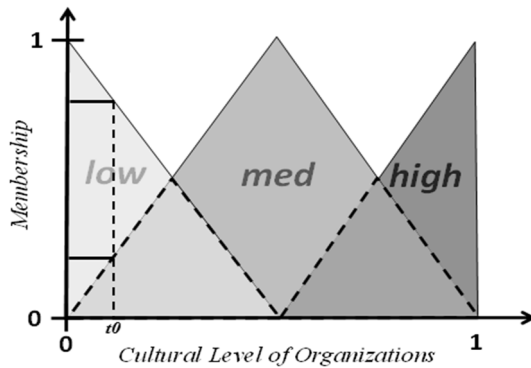
Figure 5. Membership functions denoting three fuzzy labels: Low, Medium and High.

Users then make decisions by using the fuzzy information gathered on the basis of their decision rules. These rules are a series of condition–action statements in a fuzzy model (Kosko 1994; Labib, Williams, and O'Connor 1998) that are employed in the process of making a decision, known as the inference step, to produce the fuzzy output. Different mechanisms of inference such as the Mamdani implication (Min–Max), the Larsen inference method (PROD–MAX) (Kecman 2001; Kosko 1994) and Average–Average are used in the dynamic systems context (Sabounchi et al. 2011). Figure 6 exemplifies some of these If–Then rules for two variables, level of culture and level of technology in organizations, which determines their desire for participation in the SDI development. Figure 6 represent a Min–Max inference method.
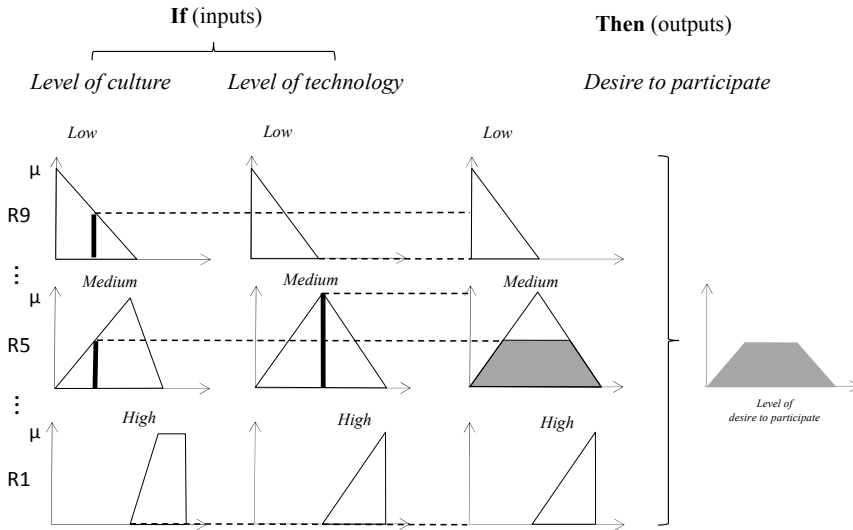
Figure 6. If-Then rules in inference step.

The fuzzy output is, in fact, the membership value to the fuzzy labels of the possible decision, as shown in the right-hand side of Figure 6. The fuzzy output should then be translated into a crisp discrete value upon which the final action must take place. The defuzzification step, the last step of a fuzzy model, is responsible for calculating the crisp value (Kosko 1994; Labib, Williams, and O'Connor 1998). Various defuzzification methods have been developed, e.g., Largest of Maximum (LOM) and the Center of Area (COA) models (Figure 7), each of which has unique advantages and disadvantages (Kecman 2001).
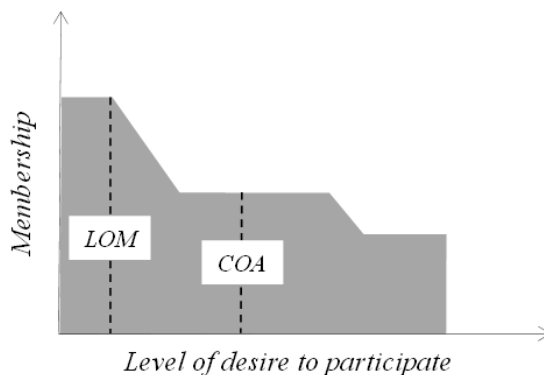
Figure 7. Largest of Maximum (LOM) and Center of Area (COA) deffuzification methods.

Studies have been conducted to represent uncertainties of the qualitative variables in the SD technique by employing fuzzy logic. Campuzano, Mula, and Peidro (2010) utilized fuzzy logic for modeling a customer–producer–employment issue. Liu, Triantis, and Sarangi (2011) also used fuzzy logic to model a combination of two linguistic variables, delivery timeliness and customer service, in a sales and service model. The effect of alternative defuzzification methods on the dynamic behavior of a model was then investigated by Sabounchi et al. (2011). They concluded that the counterintuitive behaviors of the fuzzified model were caused by discontinuous inference methods and inconsistent rules.

Various studies have highlighted that employing fuzzy logic in a dynamic system is context dependent in terms of defining rules and utilizing different methods of inference and defuzzification (Kunsch and Fortemps 2004; Kunsch and Springael 2008; Liu, Triantis, and Sarangi 2011; Mutingi and Mbohwa 2012). Polat and Bozdag (2002) and Mutingi and Mbohwa (2012) argued that differences between a crisp and fuzzy model of a dynamic system may vary across contexts.

Paper II in this PhD thesis strives to refine the SMSDI by using fuzzy logic. Four different fuzzy models are then investigated to select an appropriate model to represent the dynamic behavior of SDI development. In that paper, two linguistic variables are fuzzified, and the fuzzy models under investigation are tested by modeling the joint effect of these two variables.

3 Geospatial data integration

While an SDI provides a platform to facilitate spatial data access and sharing between various data producers and users, the importance of developing appropriate methods for integrating spatial data is increasing. Although the SDI framework seeks higher interoperability among data resources by collaboratively developing appropriate standards, this collaborative process may require a considerable amount of time to be successful. This can be realized by the increasing number of studies conducted to develop algorithms for matching and integrating spatial datasets in the past two decades.

Data integration can have different objectives such as quality improvement and dataset maintenance. However, they all share the first step, which is to find the homologous objects in two or more datasets; this process is known as object matching, instance matching, or feature matching. The corresponding objects can be detected on the basis of their similarities in geometry, topology, attributes, and semantics or combinations of these parameters.

The second part of this chapter focuses on data integration and the development of efficient and effective algorithms for matching two road network datasets from authorities and a VGI project as a case study. The road network is a major dataset that is used as the topographic data in many projects and applications. A road network dataset is a linear dataset that connects points of interest to each other if a road exists between them. Road objects are classified on the basis of their width, road cover, number of lanes, or type of transportation used. Integrating authoritative datasets with VGIs that are rich in information can enrich authoritative datasets; the quality of VGIs can also be evaluated or improved in this integration process. Using the geometry property of road network features as the most prominent property can divide the matching algorithms into two main categories: segment-based and node-based algorithms. Because these algorithms are localized on nodes or segments, their view is restricted on the structures in which the nodes and segments under matching are located. Hence, matching results can be erroneous in complex structures in which the algorithm must include the context to find the best matching pairs. Pattern detection is an approach for detecting such complex structures and can enhance the localized view of the aforementioned matching algorithms. Using data models such as RDF, which explicitly describes the representational information, is an additional solution. This

methodology is not solely dependent on the geometry of the data; it also employs the provided descriptions and relationships among features, which are key for matching complex structures connected to each other.

3.1 Segment-based algorithm

The segment-based algorithm focuses on two levels: segment and feature levels. Segments are defined as the links between vertices, or a vertex and a node; features are the links between two nodes, which may have none to many vertices in between. That is, one or more segments can compose a feature. This type of algorithm investigates the segments and features. Some algorithms explore only the similarities between segments, whereas others focus only on the features. Moreover, some algorithms begin matching in the segment level and continue the matching process in the feature level by recomposing the segments to rebuild features (Will 2014; Koukoletsos, Haklay, and Ellul 2012).

A segment-based matching algorithm begins by buffering around each segment in the reference dataset in order to find candidate segments in the target dataset located within the buffer. Other geometrical and attributive measures are used to decrease the number of candidates until the correct corresponding segments are found. After the segment level, the feature-level matching begins by first recomposing features from the segments. The matching information of the segments is also transferred to the features. A reference feature is then considered to be matched if more than half of its length is matched in the segment level. The name similarity of the non-matched target feature with the reference features can be used for matching features as well.

Walter and Fritsch (1999) developed one of the first segment-based algorithms, which is based on statistical investigation. They mapped the matching problem onto a communication system and used measures derived from the information theory for finding the optimal solution. Their approach has five steps. After preprocessing the data to reduce the global error, in the second step, they created an initial list of candidates using buffers around the reference dataset. Afterward, they removed the unlikely candidates by utilizing geometric constraints in the third step. In this step, they used manual matching in order to find the minimum and maximum thresholds for the similarity measures. In step four, they employed a merit function based on the communication system and information theory to evaluate the matching pairs detected in the previous steps. Finally, they detected a unique combination of matching pairs as the final solution for the matching in the fifth step.

Ludwig, Voss, and Krause-Traudes (2011) then proposed a segment-based method built upon the work of Walter and Fritsch (1999) and Devogele, Parent, and Spaccapietra (1998) and compared the objects using geometries and thematic attributes. After data preparation, data model matching between object classes and their attributes was executed. Then, object level matching was performed by detecting the preliminary candidates using three buffer distances. The candidates were then reduced by calculating the similarities and ranking the possible matching pairs based on length, name, and category attributes. A geographic information system (GIS) was used to visualize the ranking of the matches as the post processing, which led to the removal of some rankings from the results. Finally, the results were statistically evaluated in terms of positional difference, relative completeness of attributes, difference in speed limit, and completeness of objects.

Yang, Zhang, and Luan (2013) used a probabilistic matrix for developing a link-based matching algorithm. The probabilistic matrix, which is computed on the basis of shape dissimilarities of the candidates, is then iteratively updated by the relative compatibility coefficient of the neighboring candidate pairs. This process continues until the matrix is globally consistent. By using the probabilistic matrix, 1:1 and 1:N matching pairs can be found; another matching procedure is coupled with this algorithm to detect the M:N matching pairs. The entire matching algorithm is direction independent; that is, it can match dataset1 to dataset2 and vice versa.

Yang et al. (2014) then combined hierarchical strokes with the probabilistic relaxation method. In the new algorithm, strokes rather than links were used as the basis for matching. A stroke is a sequence of connected links that meets a certain condition such as good continuity. A set of links has good continuity if a pair of connected links does not have an angle difference above a threshold. Using strokes retains the connectivity of the road network. The hierarchical strokes are then based on the hierarchy existing among different road types. Their matching algorithm begins with matching high-grade road types, and the matched strokes in this step are used as stable references for the next layer of road with a lower grade. The iterative probability relaxation method is used for matching the strokes in each layer of road.

Schäfers and Lipeck (2014) proposed a matching algorithm considering the weighted similarity index by using geometric, semantic, and topologic similarity measures. They used a greedy approach along with optimization methods to produce minimum candidates. In this method, the run time is substantially improved while producing high-quality matching results. They imposed must-match and cannot-match constraints on the candidates and used a greedy approach to achieve the local optimization. An example for the constraint is a blacklist for

object types, i.e., an object with a field track type cannot match a highway object. Then, similarity measures such as *Starting/EndPointSimilarity* and *lengthSimilairty* as geometric similarity measures, *NameSimilarity* as an attributive similarity measure, and *DirectNeighbourhoodSimilarity* as a relational similarity measure are calculated. The relational similarity measure utilizes currently confirmed matches connected to the object under investigation to calculate its similarity. If the neighboring objects are already confirmed as matched pairs, it is more likely that the object in question has a valid match if a match is possible. This procedure produces 1:1 matching pairs; therefore, the authors added an N:M matching section where objects are aggregated, and same similarity measures were then used for the aggregated objects to find their best matching pair.

Koukoletsos, Haklay, and Ellul (2012) developed another segment-based algorithm in order to evaluate the OpenStreetMap dataset by matching it with the Integrated Transport Network (ITN) dataset from the Ordnance Survey. They divided the datasets into 1 km^2 tiles to improve the performance and to better represent the heterogeneities in the datasets. The algorithm begins at the segment level and breaks features into segments. Then, the candidate segments extracted from the buffer around the reference feature are scrutinized in terms of their geometric and attributive similarities. The matching process continues in the feature level after merging segments and adding their matching information to the features. In this level, the features are matched on the basis of the portion of their length matched in the segment level. Will (2014) extended this algorithm by adding a final check to match non-matched features according to the matched information in both datasets.

3.2 Node-based algorithm

Node-based algorithms initiate the matching process by comparing the nodes in two datasets. The nodes, which are located at a particular distance from the reference node, are considered as candidate nodes. The topological information of nodes in terms of links connected to them is checked if there is more than one candidate for a reference node. This topological relationship is known as a composition relationship between a node and its connected lines. More similarity measures may be employed if the topological comparison is not decisive. The links connected to these matched nodes are then compared, and the links with high similarities are considered as corresponding links. If two links under investigation have names, their names are used in the comparison. This similarity measure can also be used in the second step, in which the corresponding nodes are found if the other measures are not conclusive.

30

Java Conflation Suite (JCS) (Solutions 2016) is an open-source Java library that automatically matches two road-network datasets based on the geometric properties of the features. The JCS uses a node-based approach in which the best matching node of the other dataset is detected if they are located within a maximum searching distance around each node of the reference dataset. The Hausdorff distance, edge length, and angle measurements are then utilized to match the links. Links having different lengths are split and rechecked. JCS then interactively transfers the attributes between matched pairs and adds the missing links. Moreover, it provides one final network as the result of the conflation.

The JCS algorithm was adapted by Stigmar (2005) to integrate external route data with the road network data provided by a national mapping agency. She increased the matching quality by adding three specifically tailored extensions to JCS for the involved datasets. The route data was first imported to the computation environment by using Extensible Stylesheet Language Transformations (XSLT) (Kay 2007), which is an eXtended Markup Language (XML)-based transformation language that translates the input data into the desired output according to a given schema. Afterward, the route should be matched to the road network features that have different geometry and Level of Detail. This step involves JCS by first matching nodes and then edges connected to the matched nodes. However, the JCS algorithm needs to be extended because there may not be enough nodes to begin owing to different segment representations of the datasets. Therefore, the first extension, the *Merge* extension, is added before JCS is used. This extension is performed on the road network data to simplify features and remove the redundant nodes. The *Topology* and *Buffer* extensions are then added after the JCS algorithm is executed. These extensions match the unmatched segments. The *Topology* extension checks the neighboring segment of the unmatched segment, whereas the *Buffer* extension detects the segments located in the buffer around the unmatched segment to find the potential matches.

Volz (2006) also developed an iterative node-based matching algorithm coupled with the enhanced segment-based approach. This algorithm begins with a rubber-sheeting algorithm used to eliminate the geometric distortions between the datasets. Then, nodes with high likelihood of having correspondences constitute the seed nodes as starting points of the matching process, which is followed by detection of the homologous segments in order to find 1:1 matches. If there are no 1:1 matches, the enhanced segment matching is initiated to find 1:2 matches. These steps are repeated until the relaxing constraints are met. This algorithm is designed to determine the degree of inconsistency of the multiple representations as explicit relations between the matched features of two datasets.

Mustière and Devogele (2008) suggested a node-based matching routine accompanied by segment-based matching. They used this several-step process to

find the homologous features in datasets with different levels of detail. This algorithm finds one to many links between networks by using geometry, attributes, and topologic information. This algorithm is based on two principles. First, node- and segment-based matching approaches complement each other through the matching process. Second, the matching algorithm follows the roughing-out and focusing approach. That is, it first rough out the network by detecting possible matching candidates and then focuses on the candidate list. To use this algorithm, the network needs to be transformed into a common graph structure. Then, pre-matching between nodes and arcs is performed by finding the closest candidates. The selection principle is designed to reduce the sensitivity of the algorithm to the thresholds. Afterward, the most complicated step is finding the final correct matching nodes by using the preliminary matching of the previous steps. In this step, turning and direction criteria are also considered. Finally, this algorithm is able to self-evaluate its results by comparing the topological organization of networks.

Zhang and Meng (2008) suggested a network-based matching algorithm. This algorithm is a delimited stroke-oriented (DSO) approach that benefits from contextual information for matching road networks. It first creates an index that carries the relationships between connected objects. In the next step, the delimited strokes (DS) are built on the basis of the good continuity principle. The matching of DSs then begins based on the node proximity calculated in terms of their Euclidean distances, topological discrepancies, and angle differences among the links connected to the nodes. It continues by measuring their geometrical similarities based on six different criteria. If the lengths of two strokes are different, then iteratively more strokes are added to the shorter DS until their terminating points are located in each other's searching area. If no candidate is found for the starting node of a DS, the end node of the DS is examined. The network-based matching follows the stroke-matching step by continuing to check the strokes that are connected to the currently matched DS. The Matching Growing step occurs after network matching to match the links that are similar to part of a DS.

3.3 Pattern detection

Most of the aforementioned studies on matching algorithms produced promising results; however, they include mismatches in addressing complex structures in the road networks. A complex structure is composed of several links and nodes that together represent a road feature such as a dual carriageway, roundabout, or crossroads. These features exhibit patterns that are regularly repeated through the

road network. The patterns are composed of objects in a map that have properties, such as shapes, orientation, or functionality (Mackaness and Edwards 2002; Touya 2010).

In different geospatial data applications such as road network matching, generalizations, and multi-representational databases, these groups of objects need to be treated differently based on their certain characteristics. For instance, in generalization, a dual carriageway depicted with two (almost) parallel lines should be generalized into one line in a smaller-scale map. The cardinality between corresponding features in road network matching would also depend on such characteristics of, i.e., a dual carriageway. Therefore, detecting these patterns can be beneficial for determining their associated characteristics for appropriate treatment.

Pattern detection methods have been extensively used in the generalization community (Brassel and Weibel 1988; Mackaness and Edwards 2002; Heinzle, Anders, and Sester 2005; Touya 2010; Weiss and Weibel 2014; Savino, Rumor, and Lissandron 2009). Mackaness and Edwards (2002) suggested a combination of spatial clustering and graph-based techniques for detecting road junctions. They argued that identification of a junction is a scale-dependent problem, e.g., a collection of roads in a town can be viewed as a junction in a very small scale that should be depicted with a single point. Therefore, their definition of a junction in a graph is a dense cluster of nodes with degrees of three or more that were detected by using the spatial clustering model.

Savino, Rumor, and Lissandron (2009) suggested an approach in which road junctions are detected by analyzing the cycles in the road network and applying morphological analysis. This method allows classification of different junctions and generalization of the junctions in an ad-hoc manner. The authors first grouped the junctions into simple and complex junctions. The complex junctions that were then detected by cycles in the graph representation of the road were generalized. The complex junctions were further categorized into four different types based on the following taxonomy: roundabout, Δ-crossroad, Δ-junction, and paired Δ-junction (Savino, Rumor, and Lissandron (2009). Complex structure detection is performed in two steps. First simple T-intersections and simple junctions that should not be generalized are detected. In the second step, the redundant links in a junction that render the junction complex are detected. These redundant links create a cycle or a road loop. Because the nature of the road graph is highly cyclic, thresholds are set to exclude the cycles in which the areas and perimeters are beyond the threshold. Moreover, a building layer is used to exclude the roads around the blocks. The roundabouts are then detected by using the ratio of area and perimeter of the loops. The more complex junctions are also extracted by

merging the adjoining loops and reconstructing roads by using grouping principles such as the straightest road.

Touya (2010) selected road network features in the context of spatial database generalization by first detecting complex structures such as roundabouts and highway interchanges by using pattern detection. In this method, the datasets are enriched with explicit geographic structures that can help to preserve the significant structures through the generalization process. First, the crossroads are classified according to their shapes as *T-node*, *y-node*, *fork*, *star*, and *cross-shaped* (CRS). These intersections can help to detect more complex structures and to typify the structures. Roundabouts, another pattern in the road network, are also detected by using the compactness measure for a polygon:

$$Compactness = \frac{4\pi \times Area}{Perimeter^2}, \tag{Eq. 1}$$

where Area and Perimeter are the area and perimeter of the polygon under investigation.

Dual carriageways were also found by checking the shape of the polygons that are narrow and long by using the compactness (equation 1), convexity (equation 2), and elongation (equation 3) indices. Convexity and elongation are defined as

$$Convexity = \frac{Area}{HullArea}, \tag{Eq. 2}$$

$$Elongation = \frac{L}{W}, \tag{Eq. 3}$$

where *HullArea* is the area of the convex hull polygon for a given polygon, and *L* and *W* are the length and width of the minimum bounding box around the given polygon, respectively.

Touya (2010) detected highway interchanges by finding clusters of the *y-node* and *fork nodes* in the road network. The road segments located in the buffered area of the convex hull around these nodes were considered as the highway interchange features.

Few researchers in the field of road network data matching have noted the importance of enriching datasets by detecting complex structures based on their patterns (Zhang, Meng, and Bobrich 2010; Yang, Luan, and Zhang 2014). Yang, Luan, and Zhang (2014) employed methods developed in their previous papers for detecting the overall grid-like pattern of a road network (Yang, Luan, and Li 2010) and extracting complex structures in order to improve the good continuity in building strokes (Yang, Luan, and Li 2011). Yang, Luan, and Li (2010) attempted to detect a grid-like pattern in a road network by generating polygons from node-edge topology based on their relationships and a set of parameters. The study of

Yang, Luan, and Li (2011), which is more aligned to our research interest, then suggested the detection and removal of complex structures such as divided highways and roundabouts so that good continuity of the strokes in the road network can be maintained. They detected dual carriageways by using a growing buffer around each segment of the road network to find the candidate segments. They then used a heuristic tracking method to label the candidates in different groups of dual carriageways based on the good continuity principle. The algorithm examines all of the pairs and their connected segments to find the longest set of pairs as the dual carriageway. To identify the complex junctions, the authors proposed using the density-based clustering method by finding the neighboring intersection within a search area (network distance) of a given intersection.

Zhang, Meng, and Bobrich (2010) extended the road matching algorithm by Zhang and Meng (2008) to utilize the structural information. This algorithm, which is based on the delimited strokes, detects the complex structures before matching them. Roundabouts are extracted by generating isolated strokes; dual carriageways are found if two closely located polylines with similar geometric properties do not intersect. Then, each complex structure is assigned to an appropriate matching strategy. These strategies are integrated in a normal matching process. That is, if a dual carriageway in a reference dataset is unable to find its corresponding object in the target dataset, it will be considered as a normal object and will be matched by using the normal matching process.

Paper III of this PhD thesis suggests the use of a pattern detection method and a dedicated matching process for roundabouts. In this method, the algorithm begins by matching roundabouts with more contextual information and produces strong tie points between two road network datasets for matching other features. An extended node-based algorithm is also presented that employs complicated topological, geometrical, and attributive measures to match two road networks.

3.4 Ontology

Whereas SDI attempts to facilitate data management and access through distributed data resources, spatial data integration studies strive to match the heterogeneous datasets by using geometric, topologic, and attributive information to increase the interoperability among these resources. The geometric property of spatial data is considered as the main identifier for spatial features to be used in matching algorithms; however the geometry can vary owing to differences in scale, data producers with different perceptions of feature depiction, and updates. It could be beneficial to create semantic descriptions about the features and their associated shapes in a more general level before investigating the geometries.

Such information renders the matching algorithms more flexible in dealing with complex structures.

RDF is a standard descriptive data model used in the Semantic Web and Linked Data to build ontologies for richer data integration. An ontology plays an essential role in knowledge sharing (Hart and Dolbear 2013) because it explicitly defines the shared conceptualizations (Gruber 1993) in a domain by using formal ontology languages such as RDF schema or Web Ontology Language (OWL). Ontologies are also considered as conceptual data schema (Baglioni et al. 2007), which improves the interoperability levels.

There are two groups of ontologies which are mostly developed by experts and scientists: domain ontologies and application ontologies. Domain ontologies contain mainly terms in a general area of expertise, whereas application ontologies describe the terms used in a specific application (Hart and Dolbear 2013). Nevertheless, all types of ontologies are built of a set of statements known as assertions or axioms. These axioms are categorized into three groups: terminological box (TBox), relational box (RBox), and assertional box (ABox) (Rudolph 2011; Krötzsch, Simancik, and Horrocks 2012). The assertions for defining classes and relationships among them are respectively located in TBox and RBox. Some researchers consider RBox as a subdivision of TBox. ABox includes the assertions of concept instances and the relationships among them.

Several studies and projects have been launched for developing and designing ontologies for spatial data as a solution for addressing the semantic discrepancies among datasets. The LinkedGeoData (LGD) project (Auer, Lehmann, and Hellmann 2009) attempts to generate a rich geographic dataset that is integrated and interlinked under the Semantic Web concept. The LGD systematically maps OSM data to an RDF by using the key-value structures of OSM tags and then interlinks these data with DBpedia, GeoNames, and the ontology provided by the United Nations Food and Agriculture Organization (FAO) (Stadler et al. 2012). First, a Uniform Resource Identifier (URI) is created for nodes and ways in OSM data, and their tags are then mapped to specific properties and objects in the RDF. Each tag is mapped in isolation, e.g., an object with the tag "amenity=school" becomes an instance of the class "school." Therefore, the developed RDF structure is similar to the OSM structure.

Integration of the OSM vector dataset with DBpedia was also attempted by Ballatore and Bertolotto (2011) to connect a given geo-location to ontological concepts and entities. The authors claim that their work is different from LGD in terms of focusing on the user interests and considering the map scale in their processes. Their system receives a spatial query from a user with a specific location, search area distance, map scale, and number of OSM objects for retrieval. The *Semantic Service*, which is the core processor of this system, first

retrieves objects in the defined radius that considers the scale. In step 2, the object IDs are mapped to the DBpedia through LGD. The key words associated with the entity in DBpedia are then extracted to find useful semantic information in step 3. By using these keywords, a DBpedia lookup service is used to return the semantically relevant resources (URIs) in step 4. These resources are scrutinized in step 5 to ensure their validity by checking their geographic proximities and tag matching. The semantic information is further processed to reach parent classes in step 6. Finally, in step7, the results can be stored in XML or visualized for the user.

During the past decade, different organizations including Open Geospatial Consortium (OGC), W3C, triple-store vendors, and researchers have attempted to develop strategies for representing and querying spatial data in an RDF data model. OGC suggested the first standardized GeoSPARQL language as a platform for developing spatial data ontologies by proposing a vocabulary for representing geospatial data in RDF. GeoSPARQL also extends the SPARQL query language with query predicates and functions for processing geospatial data (Perry and Herring 2011; Battle and Kolas 2011). In a parallel attempt, W3C and NeoGeo created a geospatial vocabulary as well. The W3C proposed the Basic RDF Geo vocabulary (W3C Semantic Web Interest Group 2003). The W3C Geo vocabulary was then updated by the Geospatial Incubator Group (GeoXG) (Lieberman, Singh, and Goad 2007). A similar project was launched in 2009 by the NeoGeo community, which was taken over by GeoVocab.org (Salas et al. 2011).

3.5 Ontology in data integration

Ontologies are also used in geospatial data integration (Uitermark et al. 1999; Kavouras and Kokla 2000; Du et al. 2012). Uitermark et al. (1999) utilized ontologies in order to integrate geospatial datasets. They argued that data interoperability is a communication problem that requires a language such as ontology, which is built upon shared concepts. They detected semantically similar concepts by defining abstraction rules between a domain ontology and the application ontologies of datasets. In this method, application ontologies are first mapped to the domain ontologies with two relations: equivalent classes aggregated classes. Two equivalent classes are semantically similar in both domain and application ontologies, whereas the aggregated class in an application ontology is built of two or more classes in the domain ontology. According to these relations, two concepts in the two application ontologies are considered to be semantically equal if they refer to the same concept in the domain ontology. Concept A in application ontology 1 is semantically *related* to concept B in application ontology

2 if concept A refers to a concept in the domain ontology that is a subclass/superclass of the concepts referred to concept B in the domain ontology. Finally, concept A in application ontology 1 is semantically *relevant* to concept B in application ontology 2 if the concept referred to in the domain ontology is an aggregated class for concept A. All of these relations are manually produced and used for finding the corresponding objects between two dataset. Therefore, two object having similar semantics and sharing the same location are considered as matched pairs. The overlay function is used for determining objects with the same location.

Kavouras and Kokla (2000) then developed a methodology based on Formal Concept Analysis (Wille 1992) to semantically integrate heterogeneous geographic databases. They fused the categorizations in databases to integrate spatial data. Their work is based on two steps: Semantic Factoring and Concept Lattices. In the first step, all of the concepts in the databases are analyzed and decomposed into fundamental classes. The equivalent and overlap relations between classes are also specified. According to the decomposition rule, the overlapping classes are then split into disjoint classes to create simpler concepts. The second step utilizes the basic notions of the Formal Concept Analysis to combine the decomposed simple classes into Concept Lattices, which are the integrated structures of categorizations in different databases that determine the association and interaction among those databases.

Du et al. (2012) also developed an ontology-based matching algorithm for road network datasets. They first converted datasets into ontologies based on their given data models. They represented the road vector data as a graph in the ontologies composed of edges and vertices. Therefore, the produced ontologies have two main classes of Edge and Vertex with two respective properties connecting them: *hasVertex* and *isVertexOf*. The information about the road features are then stored as properties of the *NamedIndividuals* (instances in an ontology) according to the data schema of each dataset. The *NamedIndividuals* are created on the basis of the names used as identifiers to detect the corresponding instances between two ontologies. Afterward, the ontologies are merged, and a new ontology is created that carries information from both datasets. In the next step, the algorithm detects topological and geometrical inconsistences in the created ontology and allows users to interactively fix the inconsistencies.

Several projects and studies have attempted to develop ontologies that include road features. The Reasoning on the Web with Rules and Semantics (REWERSE) project proposes an ontology for a transportation network that mainly represents the public transportation network (Lorenz, Ohlbach, and Yang 2005). Codescu et al. (2011) developed an ontology for OSM datasets known as OSMonto, which also includes concepts used in road network. The OSMonto is based on the key-

tag structure of data in OSM and facilitates representation and study of the hierarchical structure of the OSM tags. This method enables connection of the tags to other ontologies in order to enrich their semantics. Finally, OSMonto helps to unify concepts shared among different tags. This ontology, however, acts as metadata and does not include instances of spatial features.

A novel data structure in an RDF data model is presented in Paper IV of this PhD thesis, which includes new representational concepts for a road network. These concepts explicitly define the representations of features and their relationships. Using such a structure can facilitate the matching of complex structures in road networks and can improve the matching results.

3.6 Volunteered Geographic Information

The emergence of Web 2.0 has enabled users to interact with Websites, which fosters a new generation of data collection. VGI as part of data produced on the Internet is cost-efficiently produced and frequently updated by volunteer contributors (Du et al. 2013). This has led to the emergence of several projects based on voluntarily gathered spatial data such as that offered on OpenStreetMap (OSM), Google Map Maker, Wikimapia, and other Websites. Some governments have also attempted to adopt the VGI approach to reflect the local needs or problems in collected spatial information (Ghose 2003).

The OSM project is by far the most popular VGI project. OSM distributes its voluntarily collected data under the Open Database License (ODBL) (Wiki 2016). The initial idea of the OSM project was to collect the road network data of the United Kingdom. It soon developed into a worldwide project with an increasing number of users (Figure 8) collecting data in OSM. The data voluntarily collected by the users also began to include all types of features such as roads, points of interest, buildings, and land cover. Soon afterward, the community began to develop open tools for collecting, uploading, downloading, and visualizing the data, which led into an increasingly user-friendly environment. These improvements have attracted more people to OSM, including those with no background in the geographic information field.
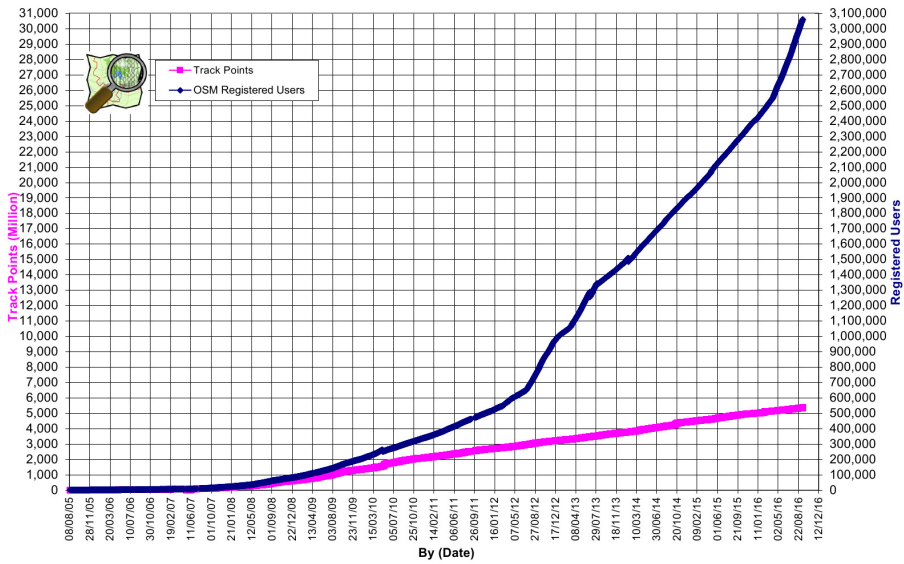
Figure 8. OSM contributors and GPS points uploaded from 2005 until 2016 (from OSM Wiki).

Currently, several data collection methods are available in OSM. The most common is Global Positioning System (GPS) survey. Digitization of orthophotos was introduced after companies such as Yahoo and Bing provided OSM with their own orthophotos. The walking papers and field papers methods are additional data collection methods used to produce local and attribute maps without employing GPS. The contributor simply prints out part of the OSM map and manually adds the attributes. Afterward, the annotated map can be scanned and uploaded to the OSM to apply the changes. The actual changes to the OSM data can be made through available OSM editors such as iD (iD 2016), Potlatch 2 (Patlatch2 2016), and JOSM (JOSM 2016). The first two editors are online editors available on the OSM Website. These editors target new users by providing basic functionalities for adding and modifying the features. On the contrary, JOSM, which has extended functionalities, is designed for more advanced users. This editor is offline and needs to download data from OSM and upload them back to OSM after manipulation.

OSM has a simple data model composed of three main elements: node, way, and relation. Nodes and ways represent the geometry of features in OSM. A node has an ID and is defined by a pair of coordinates to represent a specific point on Earth, whereas a way is an ordered list of nodes that represents linear features. Nodes and

40

ways can relate to each other and comprise a larger entity. This entity is built by using the relation element in OSM in which the involved nodes and ways are members, each having a unique role. The relation element is optional and it does not exist for all features. All of these elements can have several attributes using a key-value tagging structure. In this regard OSM data is considered to be structured compared with the Wikimapia system, which collects unstructured content from contributors in the form of free-text descriptions or links.

However, the definition of structured data also varies among fields. For example, in the geographic information (GI) field, the spaghetti data model for geometries is considered to be unstructured, whereas other research (Rinner and Fast 2015) considers OSM data, which has a spaghetti model, as structured. Moreover, the data can be structured in different ways such as data layer, schema, and feature- or instance-level structures. The data layer structure is mostly popular in file-based systems in which data are divided and stored in different layers, such as road networks, buildings, and land use and land cover maps. In databases, the data are organized as entities/classes and the relationships between them. OSM data may have data layer structure because it has different layers. The data schema is also considered in OSMonto, but feature- or instance-level structures do not exist in OSM data. The feature or instance level structure is about how different instances in a layer are related to higher-level concepts and to each other. For example, a street in a road layer is constructed of a set of different sections, or a roundabout connects several roads.

The relation element in OSM can be used to build the instance-level structure; however, the relationships among features are optional and may not exist for all features. In fact, most VGI projects do not bind their users to such a data structure in order to keep the data collection process simple. In this regard, the feasibility of employing a simple supervisory mechanism in order to collect well-structured VGI is investigates in Paper V of this PhD thesis. This paper uses a step-wise method in order to structure data in the instance level based on a data structure proposed in paper IV. The volunteer participants then used the developed application to collect data. The collected data along with a questionnaire were used to evaluate the employed mechanism.

4 Summary of papers

4.1 Paper I

Spatial Data Infrastructure (SDI) as a platform to be built through a spatial data community faces many challenges owing to the high complexity of SDI, lack of knowledge and experience, and limited insight in the benefits. Various models have been developed to improve the understanding of the SDI nature and its complexity. The state-of-the-art simulation model developed by Mansourian and Abdolmajidi (2011) is able to simulate the progress of SDI development regarding the employed policies and can provide policy-makers with insight into the strengths and weaknesses of SDI development policies. Adapting this simulation model by using the communities of practice concept has facilitated development of a methodology for SDI planning in Tanzania in this study.

A community consisting of experts from stakeholder organizations gathered to develop an SDI plan, where they improved their understanding of SDI by exchanging their knowledge and discussing their ideas. They also learned about the future effects of their current plans and decisions for developing SDI by deploying their policies into the simulation model of SDI development (SMSDI). This simulation model was developed by using the system dynamics (SD) technique, which can present the SDI complexity in terms of its major influencing factors and their timely interactions. The SMSDI can simulate the progress of implementing SDI over time, which gives the planning community an approximate image of the future. By using the SMSDI, SDI stakeholders could form a consensus on the optimum plan by refining and improving the model.

This model includes the components and policies that are essential for planning successful SDI implementation in Tanzania and can help to gain political support. The model made it possible to present the challenges related to the current status of the national SDI to the executive organizations' experts (EEO), who are the SDI Steering Committee representatives in this study. The results also persuaded these officials to support the model for the development of a national SDI in the country.

4.1.1 Conclusions

Using the SD technique in the community of practice context facilitates and improves SDI planning because the stakeholders learn while sharing their knowledge. The different educational and scientific backgrounds of EEO lead to different levels of awareness and understanding of SDI, which makes it very challenging to involve all these people in the planning process. The community of practice enables EEO members to take part in the SDI planning and helps to improve their knowledge of SDI.

Therefore, the suggested integrated approach can increase the knowledge of organizations about SDI and its complex planning process through indirect learning, which motivates them to participate in SDI planning. It was observed that the SMSDI model is more tangible for decision-makers because they were able to implement their policies and observe the results in terms of SDI development over time. The model received high credit from the decision makers because the majority of the SDI stakeholders agreed upon the models. This research taught us the following valuable lessons:

(1) The SD technique can respond to the challenges in modeling SDI as a complex system.

(2) The integrated approach can to some extent motivate the decision-maker to support SDI development in Africa.

(3) We can argue the feasibility of applying this methodology for SDI planning in countries similar to Tanzania.

4.2 Paper II

The system dynamics (SD) technique has been used to simulate the development of spatial data infrastructures (SDI) in order to provide insight into the future progress of the system regarding the employed policy. An SDI is a system composed of various dynamically interacting quantitative and qualitative (linguistic) variables, which creates a collaborative environment to facilitate the management of and access to the spatial data at different political and administrative levels. SD is able to bring together the qualitative and quantitative factors in system modeling; however, to incorporate the uncertainty of the linguistic variables and their joint effects in an SDI development model more properly, fuzzy logic may be integrated with SD. Such integration can be used to interpret the current status of various qualitative factors in a manner that is more readily understood and applied for the key players in a system. With this approach

in mind, we suggest redefining the modeling of the joint effect of linguistic variables in the simulation model of SMSDI using fuzzy logic because the model involves several linguistic variables.

Numerous fuzzy models exist, all with advantages and disadvantages. Considering the unavoidable inconsistencies in rule definitions of SDI and the incompatibility of fuzzy models with dynamic systems, a proper fuzzy model is required that can model the dynamic behavior of SDI in SMSDI. Therefore, this study aims to investigate different fuzzy models and their suitability for modeling SDIs. In this regard, different combination of two inference and two defuzzification methods were used to fuzzify the joint effect of two variables in an existing SDI model.

The joint effect of two linguistic factors in the SMSDI, *level of culture* and *level of technology*, were modeled by using four fuzzy models. These four models are different combinations of two inference methods, Max–Min (Mamdani) and Average–Average, and two defuzzification methods, Largest of Maximum (LOM) and Center of Area (COA). A new fuzzy structure was added to SMSDI to model the joint effect of the two influencing factors. Afterward, different fuzzy models were applied and evaluated by behavior reproduction and sensitivity analysis tests.

4.2.1 Conclusions

The results showed that the combination of COA defuzzification with Average–Average inference reflects the dynamic behavior of SDI development better than the other three models. Because of the averaging nature of the COA defuzzification, the COA–Ave–Ave model still bears some counterintuitive behavior at extreme points. This necessitates investigation of more fuzzy models in future studies.

The two main reasons for the counterintuitive behavior of other three models are 1) discontinuity owing to use of the maximum or minimum values in inference and defuzzification steps and 2) inconsistency in defining the rules. Rule inconsistency is a context-dependent issue. That is, inconsistency could be justifiable in the SDI development context. This problem may be also approached by using more fuzzy labels such as *very low*, *low*, *medium*, *high*, and *very high* for the fuzzy inputs and output.

4.3 Paper III

Spatial Data Infrastructure (SDI) provides a platform to facilitate the spatial data sharing and access among users and producers. It also opens a new door to the distributed data voluntarily collected on the Internet. Over the past decade, the amount of volunteered geographic information (VGI) has increased rapidly. Because these spatial resources are collected through crowd sourcing in which amateur and expert participators are involved in the data collection, the data quality varies. In this regard, the VGI data quality and methods for assessing it have been the main topic of many research studies. In this paper, we also evaluate the completeness of a VGI road network dataset using an authoritative dataset as a reference through an accurate and efficient network-matching algorithm.

To develop a matching algorithm, we first studied and compared the two main strategies for network matching: segment-based and node-based matching. According to the comparison results, the node-based algorithm is considerably more computationally efficient; however, the result quality of the two strategies is comparable. Therefore, a standard node-based algorithm was improved in terms of accuracy by adding topological relationship handling and pattern detection of the complicated network components. The extended node-based algorithm was utilized for matching OpenStreetMap (OSM) with the Swedish National Road Database (NVDB) in Scania, Sweden. The results show OSM completeness of 87% and 69% in the urban and rural areas of Scania, respectively. The accuracy of the matching process was as high as approximately 95%. However, investigating the results highlighted some shortcomings where the algorithm is not able to detect all of the patterns or relate two patterns in the road network context. This means that more pattern detection methods should be deployed and that the relationships between two complex structures must also be considered in the matching process.

4.3.1 Conclusions

In this study, the state-of-the-art node-based network-matching algorithm was improved to better detect corresponding nodes and links of complex structures in two road network datasets. This was performed by introducing roundabout detection and its exclusive matching process. The improved node-based algorithm displayed better accuracy; however, it lost time efficiency to some extent. We can conclude that the extended node-based algorithm is efficient and sufficiently accurate for conducting surveys of the quality of VGI road datasets in large geographic regions.

4.4 Paper IV

Many road matching studies frequently report representational and semantic differences among datasets as the two main obstacles. While knowledge about how an entity is defined can help to find similar entities in semantic applications, knowledge about how features are represented in datasets can also be key in improving the automatic matching process. Therefore, to address the representational heterogeneity, a methodology based on adding representational information to a road network ontology (RNO) is suggested in this study.

This idea was motivated as the usage of ontology has increased through development of the Semantic Web, and Resource Description Framework (RDF) has become a standard model for exchanging data on the Web. Additionally, well-structured datasets in an RDF data model can address the semantic issues and create a matching algorithm with greater flexibility in dealing with complex structures such as roundabouts, carriageways, and crossroads with different levels of detail. By using the RDF data model, implicit knowledge in the dataset such as the cardinalities of instances can also be extracted and used to accurately match the homologous instances. Moreover, the representational information stored in or inferred from the RDF data model can be efficiently queried by using SPARQL language rather than by processing geometrical, topological, and attributive information.

The proposed RDF data model was applied in two case study datasets. The first dataset, an authoritative dataset, was translated into an RDF data model. The translation was feasible because the authoritative dataset contained sufficient information for extracting the representational information required by the data model. On the contrary, the second dataset was voluntarily collected by using an editor that was tailored to collect well-structured road data based on the proposed structure. This dataset is geometrically and attributively in accordance with the OSM road network layer.

By using the reasoning capability, implicit information was extracted, and the inferred ontologies were fed into a node-based algorithm that was adapted to retrieve knowledge in the ontology by using SPARQL queries. The algorithm was able to access necessary information, process the relationships among features, and find correct corresponding links and nodes. Manual assessment of the results also showed high accuracy and high quality of matched features (nodes and links) while retaining road network consistency.

4.4.1 Conclusions

The RDF data model is able to store representational information by explicitly describing the features and their relationships. The implicit information in the ontology can be also inferred and used in the matching for finding the corresponding features. Moreover, using the RDF data model allows the matching algorithms to consider the semantic similarities among features. Ultimately, employing the RDF data model and SPARQL query language made the modified node-based matching algorithm more flexible in handling complex structures and improved the final accuracy. The matched network also retained its connectivity after matching, which enabled consistent road network integration. It can be concluded that the proposed RDF data model serves as an asset in instance-matching applications and it could be extended to other geospatial applications.

4.5 Paper V

Advancement in technology and widely available smart phones provides a great opportunity for users to become involved in data collection because they are enabled by Web 2.0 to interact with Web applications. Crowed-sourced geographic data known as Volunteered Geographic Information (VGI) is part of the data collected by different users actively or passively and has become a new source of geospatial data. To increase or maintain the voluntarily participation, VGI projects do not bind users to any specific data structure. As a result, the collected data are highly heterogeneous. This heterogeneity is an obstacle in integrating data from different resources. If VGI projects intend for their data to be used in other projects as an asset, it is reasonable for them to improve their data structure. On the contrary, better data structuring may result in increasing the complexity of data collection; hence, contributors could be lost.

This article aims to explore the feasibility of well-structured data collection in a VGI project by developing a step-wise Web application to supervise the data collection. This Web application is based on a proposed Resource Description Framework (RDF) data model that is able to encompass the representational information of road network features. A test group of users was asked to use the Web application to digitize a designated area in Lund, Sweden. Thirteen users voluntarily collected data and answered the given questionnaire. The results indicate that the users found the structure easy to follow, although increases in the data collection time shows the effect of the mandatory structure. The average data quality of 94% manifests the effectiveness of the application in structuring the

data. Moreover, it denotes the simplicity of the data structure and the supervisory process, which enabled the users to collect well-structured data.

4.5.1 Conclusions

In this study, a step-wise Web application was developed to investigate the feasibility of gathering well-structured road network data in a VGI project. This application collects data according to the structure of an RDF data model that embeds the representational information of road network features. Such a data structure can be a potential solution for dealing with representational heterogeneity in integrating datasets.

The web application was tested and the results revealed that although binding users with a mandatory structure increased the data collection time, the users still found the structure easy to follow. The high structure quality of the collected data also highlights the effectiveness of the application and simplicity of the proposed structure. Such simplicity led to a high likelihood of user employment of the application in the future. The Web application and therefore the data quality can be further improved by providing data manipulation/moderation tools so that users and other contributors can correct the errors in the data.

In summary, a step-wise application for collecting VGI can produce well-structured data while retaining its simplicity. The structured data can then facilitate future data integration in terms of addressing the representational heterogeneity. Using the RDF data model also provides a means for embedding the semantics in the data and enables involvement of VGIs in the Semantic Web and Linked Data.

5 Conclusions

Organizations would be more motivated to get involved in SDI development if they were able to trust the given policies. Moreover, policy-makers would be more confident if they could employ a tool for testing the reliability of their proposed policies. The SD technique can provide such a tool for modeling SDI development and for testing the policies to gain insight on the future progress of the SDI. However, to build the model, policymakers and organizations need to collaborate to establish a common understanding of the system under investigation. The community of practice provides the required framework for such a collaboration whereby different participants interactively learn about the SDI and help to develop a unanimously agreed upon model. In order for the policy-makers to better understand and interpret the simulation results, the simulation model should be able to represent the involved factors in the same manner as that used by policy-makers in the real world. Fuzzy logic is an approach that can bring the simulation model closer to the human perspective and decision-making process.

Agreeing upon the SDI development policies does not resolve all of the obstacles involved in effectively sharing and reusing the distributed spatial data resources. The need exists for the development of methodologies that can help to properly integrate the heterogeneous spatial datasets shared in the SDI platform. The heterogeneity problem is further highlighted as VGI datasets open their way to the SDI as rich and up-to-date resources.

For data integration at the instance level, different matching algorithms have been previously developed that find the corresponding objects between two datasets based on local investigation of the geometric, topologic, and attributive information of nodes and links. Although these algorithms show promising results, involving more contextual information can increase the accuracy of matching. Pattern detection methods can add such contextual information to the matching algorithms by detecting complex structures based on their representations. Another methodology is for the data producers to embed the representational information in the data by describing the complex features and their relationships. A proper data structure in an RDF data model can be a solution for dealing with representational heterogeneity. In this methodology, providing the representational information, especially in VGIs, is key. A simple supervisory mechanism in a VGI project has

been proven effective in collecting well-structured road network data by the volunteer participants.

Considering the above discussion, this PhD thesis has successfully contributed to better modeling of SDI development in two case studies and has provided novel methodologies for improving data integration in an SDI environment.

Acknowledgement

First of all, I would like to thank Ali Mansourian, my main supervisor, and Lars Harrie, my supporting supervisor, for their great support throughout this thesis. Their encouraging comments and constructive discussions gave me the motivation to accomplish my PhD.

Special thanks to Petter Pilesjö as the director of GIS center for his amazing personality and extensive support which give everyone a sense of security. Big thanks also go to all my colleagues at the GIS center: Alex, Andreas, Finn, Karin Mohammadreza, Micael, Mitch, Olive and Roger, for creating such a wonderful atmosphere to make everyone feel welcome.

I appreciate all help I received from the administrative staff in the Physical Geography and Ecosystem Science department. I also appreciate the technical support from Rafael and Ricardo.

I am grateful to all my fellow PhD students and my dearest friends for keeping me company through my hard times and always giving me positive energy. My appreciation also goes to Cecilia, Hakim, Jing and Mohammadreza for scrutinizing my manuscripts.

I also want to thank Erica who painted my days with colors and brought more joy and happiness to my life. Her presence is always a source of comfort for me.

Finally I want to express my deepest gratitude to my family, specially my lovely parents, who have been the light of my life. They always supported and encouraged me to step forward and to pursue my dreams.

References

Ahmad, K, and R. Al-Sayed. 2005. "Community of practice and the special language 'Ground'." In *Encyclopaedia of Knowledge Management and Community of Practice*, edited by S Clarke and E. Coakes. Hershey (PA): The Idea Group Reference.

Auer, S., J. Lehmann, and S. Hellmann. 2009. "LinkedGeoData: Adding a Spatial Dimension to the Web of Data." Review of. *Semantic Web - Iswc 2009, Proceedings* 5823:731-46.

Baglioni, Miriam, Maria Vittoria Masserotti, Chiara Renso, and Laura Spinsanti. 2007. "Building geospatial ontologies from geographical databases." In *GeoSpatial semantics*, 195-209. Springer.

Ballatore, A., and M. Bertolotto. 2011. "Semantically Enriching VGI in Support of Implicit Feedback Analysis." Review of. *Web and Wireless Geographical Information Systems* 6574:78-93.

Battle, Robert, and Dave Kolas. 2011. "Geosparql: enabling a geospatial semantic web." Review of. *Semantic Web Journal* 3 (4):355-70.

Brassel, Kurt E, and Robert Weibel. 1988. "A review and conceptual framework of automated map generalization." Review of. *International Journal of Geographical Information System* 2 (3):229-44.

Campuzano, Francisco, Josefa Mula, and David Peidro. 2010. "Fuzzy estimations and system dynamics for improving supply chains." Review of. *Fuzzy Sets and Systems* 161 (11):1530-42.

Chan, Tai On, Mary Feeney, Abas Rajabifard, and Ian.P Williamson. 2001. "The Dynamic Nature of Spatial Data Infrastructures: A Method of DescriptiveClassification." Review of. *GEOMATICA* Volume 55 (1):pages 65-72.

Chan, Tai On, and Ian P. Williamson. 1999. "A Model of the Decision Process For GIS Adoption and Diffusion In A Government Environment." Review of. *URISA Journal* Volume 11 (Number 2):Pages 7 - 16.

Clark, H. H., and S. E. Brennan. 1991. "Grounding in communication." Review of. *Perspectives on socially shared cognition* 13 (1991):127-49.

Codescu, Mihai, Gregor Horsinka, Oliver Kutz, Till Mossakowski, and Rafaela Rau. 2011. "Osmonto-an ontology of openstreetmap tags." Review of.

Cooper, Antony K., Harold Moellering, Jan Hjelmager, Petr Rapant, Tatiana Delgado, Dominique Laurent, Serena Coetzee, et al. 2013. "A spatial data infrastructure model from the computational viewpoint." Review of. *International Journal of Geographical Information Science* 27 (6):1133-51. doi: 10.1080/13658816.2012.741239.

Cooper, Antony, Harold Moellering, Tatiana Delgado, Ulrich Duren, Jan Hjelmager, Michel Huet, Petr Rapant, et al. 2007. "An Initial Model Of The Computation Viewpoint For A SDI." In.

Crompvoets, J., A. Bregt, A. Rajabifard, and I. Williamson. 2004. "Assessing the worldwide developments of national spatial data clearinghouses." Review of. *International Journal of Geographical Information Science* 18 (7):665-89.

Dai, Xin, JH Xiao, and Kang Xie. Growth of Enterprise Information Technology Application: System Dynamics Model and Empirical Evidence.

De Man, EWH. 2007. "Are spatial data infrastructures special." Review of. *Research and Theory in Advancing Spatial*.

Devogele, Thomas, Christine Parent, and Stefano Spaccapietra. 1998. "On spatial database integration." Review of. *International Journal of Geographical Information Science* 12 (4):335-52. doi: 10.1080/136588198241824.

Du, H. S., N. Alechina, M. Jackson, and G. Hart. 2013. "Matching Formal and Informal Geospatial Ontologies." Review of. *Geographic Information Science at the Heart of Europe*:155-71. doi: Doi 10.1007/978-3-319-00615-4_9.

Du, H. S., S. Anand, N. Alechina, J. Morley, G. Hart, D. Leibovici, M. Jackson, and M. Ware. 2012. "Geospatial Information Integration for Authoritative and Crowd Sourced Road Vector Data." Review of. *Transactions in Gis* 16 (4):455-76. doi: DOI 10.1111/j.1467-9671.2012.01303.x.

Dudley, R., and Chris Soderquist. 1999. "A simple example of how system dynamics modeling can clarify and improve discussin and modeification of model structure." In *The 129th annual meeting of the American fisheries society*. Charltte, NC.

Erik de Man, W. H. 2006. "Understanding SDI; complexity and institutionalization." Review of. *International Journal of Geographical Information Science* 20 (3):329-43. doi: 10.1080/13658810500399688.

Forrester, Jay Wright. 1958. "Industrial Dynamics - a Major Breakthrough for Decision Makers." Review of. *Harvard Business Review* 36 (4):37-&.

———. 1961. *Industrial dynamics*. Cambridge, Mass.: M.I.T. Press.

———. 1968. *Principles of systems : text and workbook, chapters 1 through 10*. 1 vols. Cambridge, Mass.: Wright-Allen.

———. 1969. *Urban dynamics*. Cambridge, Mass.,: M.I.T. Press.

Ghose, Rina. 2003. "Community participation, spatial knowledge production, and GIS use in inner-city revitalization." Review of. *Journal of Urban Technology* 10 (1):39-60.

Groot, Richard, and John D. McLaughlin. 2000. *Geospatial data infrastructure : concepts, cases, and good practice*, *Spatial information systems*. New York: Oxford University Press.

Gruber, T. R. 1993. "A Translation Approach to Portable Ontology Specifications." Review of. *Knowledge Acquisition* 5 (2):199-220. doi: DOI 10.1006/knac.1993.1008.

Grus, L., J. Crompvoets, and A. K. Bregt. 2010. "Spatial data infrastructures as complex adaptive systems." Review of. *International Journal of Geographical Information Science* 24 (3):439-63. doi: Doi 10.1080/13658810802687319.

Grus, Lukasz, Joep Crompvoets, and Arnold Bregt. 2006. Defining National Spatial Data Infrastructures as Complex Adaptive Systems. Paper presented at the GSDI-9 ntiago, Chile, 6-10 November

Hart, Glen, and Catherine Dolbear. 2013. *Linked data: A geographic perspective*: CRC Press.

Heinzle, F, KH Anders, and M Sester. 2005. Graph based approaches for recognition of patterns and implicit information in road networks. Paper presented at the Proc. 22nd International Cartographic Conference, A Coruna.

Hendriks, P. H. J., E. Dessers, and G. van Hootegem. 2012. "Reconsidering the definition of a spatial data infrastructure." Review of. *International Journal of Geographical Information Science* 26 (8):1479-94. doi: Doi 10.1080/13658816.2011.639301.

Hjelmager, Jan, Tatiana Delgado, Harold Moellering, A. K. Cooper, David Danko, Michel Huet, H. J. G. L. Aalders, and Alexander Martynenko. 2005. Developing a modelling for the spatial data infrastructure. Paper presented at the Mapping approaches into a changing world, 22nd cartographic conference, A Coruna, Spain, 9-16 July.

Hjelmager, Jan, Harold Moellering, Antony Cooper, Tatiana Delgado, Abbas Rajabifard, Petr Rapant, David Danko, et al. 2008. "An initial formal model for spatial data infrastructures." Review of. *International Journal of Geographical Information Science* 22 (11-12):1295-309. doi: 10.1080/13658810801909623.

iD. Accessed 23, Aug, 2016. http://ideditor.com.

JOSM. Accessed 23, Aug, 2016. https://josm.openstreetmap.de/.

Kavouras, Marinos, and Margarita Kokla. 2000. Ontology-based fusion of geographic databases. Paper presented at the Proceedings of the FIG Workshop:" Spatial Information Management-Experiences and Visions for the 21st Century", 4-7 October, Athens, Greece.

Kay, Michael. 2007. "Xsl transformations (xslt) version 2.0." Review of. *W3c recommendation* 23:52-71.

Kecman, V. 2001. *Learning and soft computing : support vector machines, neural networks, and fuzzy logic models*, *Complex adaptive systems*. Cambridge, Mass.: MIT Press.

Kosko, B. 1994. "Fuzzy-Systems as Universal Approximators." Review of. *Ieee Transactions on Computers* 43 (11):1329-33. doi: Doi 10.1109/12.324566.

Koukoletsos, T., M. Haklay, and C. Ellul. 2012. "Assessing data completeness of VGI through an automated matching procedure for linear data." Review of. *Transactions in Gis* 16 (4):477-98.

Krötzsch, Markus, Frantisek Simancik, and Ian Horrocks. 2012. "A description logic primer." Review of. *arXiv preprint arXiv:1201.4089*.

Kunsch, P., and J. Springael. 2008. "Simulation with system dynamics and fuzzy reasoning of a tax policy to reduce CO_2 emissions in the residential sector." Review of. *European Journal of Operational Research* 185 (3):1285-99. doi: DOI 10.1016/j.ejor.2006.05.048.

Kunsch, PL, and Ph Fortemps. 2004. Evaluation by fuzzy rules of multicriteria valued preferences in agent-based modelling. Paper presented at the Proceedings of MUDSM 2004 Conference, Coimbra, Portugal, September.

Labib, A. W., G. B. Williams, and R. F. O'Connor. 1998. "An intelligent maintenance model (system): an application of the analytic hierarchy process and a fuzzy logic rule-based controller." Review of. *Journal of the Operational Research Society* 49 (7):745-57. doi: DOI 10.1057/palgrave.jors.2600542.

Lave, J., and E. Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge, MA: Cambridge University Press.

Lieberman, Joshua, Raj Singh, and Chris Goad. 2007. "W3c geospatial vocabulary." Review of. *Incubator group report, W3C*.

Liu, S. Y., K. P. Triantis, and S. Sarangi. 2011. "Representing Qualitative Variables and Their Interactions With Fuzzy Logic in System Dynamics Modeling." Review of.

Systems Research and Behavioral Science 28 (3):245-63. doi: Doi 10.1002/Sres.1064.

Lorenz, Bernhard, Hans Jürgen Ohlbach, and Laibing Yang. 2005. "Ontology of transportation networks. REWERSE Deliverable A1-D4, University of Munich." Review of. *Institute for Informatics*.

Ludwig, Ina, Angi Voss, and Maike Krause-Traudes. 2011. "A Comparison of the Street Networks of Navteq and OSM in Germany." In *Advancing Geoinformation Science for a Changing World*, edited by Stan Geertman, Wolfgang Reinhardt and Fred Toppen, 65-84. Berlin, Heidelberg: Springer Berlin Heidelberg.

Mackaness, William, and Geoffrey Edwards. 2002. The importance of modelling pattern and structure in automated map generalisation. Paper presented at the Proceedings of the joint ISPRS/ICA workshop on multi-scale representations of spatial data.

Mansourian, A., and E. Abdolmajidi. 2011. "Investigating the system dynamics technique for the modeling and simulation of the development of spatial data infrastructures." Review of. *International Journal of Geographical Information Science* 25 (12):2001-23. doi: Doi 10.1080/13658816.2011.567390.

Mustière, Sébastien, and Thomas Devogele. 2008. "Matching networks with different levels of detail." Review of. *GeoInformatica* 12 (4):435-53.

Mutingi, Michael, and Charles Mbohwa. 2012. Fuzzy system dynamics simulation for manufacturing supply chain systems with uncertain demand. Paper presented at the Computers and Industrial Engineering 42.

Omran, El-Sayed E. , J. Crompvoets, and A. Bregt. 2006. "Benefits and bottlenecks for SDI development in Egypt." Review of. *GIS Development.* 2 (1):32-5.

Patlatch2. Accessed 23, Aug, 2016. http://www.openstreetmap.org/edit?editor=potlatch2.

Perry, Matthew, and John Herring. 2011. "OGC GeoSPARQL-A geographic query language for RDF data." Review of. *OGC Implementation Standard, ref: OGC*.

Polat, S., and C. E. Bozdag. 2002. "Comparison of fuzzy and crisp systems via system dynamics simulation." Review of. *European Journal of Operational Research* 138 (1):178-90. doi: Doi 10.1016/S0377-2217(01)00124-2.

Quaddus, Mohammed, and Arunee Intrapairot. 2008. System Dynamics Modelling of Information and Communication Technologies: Policy Studies in the Banking Industry in Thailand. Paper presented at the The 16th International Conference of the System Dynamics Society.

Rajabifard, A., and I. P. Williamson. 2003. "Anticipating the Cultural Aspects of Sharing for SDI Development." In *Spatial Sciences 2003 Conference*. Canberra, Australia.

Rajabifard, Abbas. 2001. SDI Hierarchy from Local to Global SDI Initiatives. Paper presented at the Open Seminar on SDI in Asia and the Pacific Region, 7th PCGIAP meeting, Tsukuba, Japan.

Rajabifard, Abbas, Mary-Ellen F. Feeney, and Ian P. Williamson. 2002. "Future directions for SDI development." Review of. *International Journal of Applied Earth Observation and Geoinformation* 4 (1):11-22. doi: http://dx.doi.org/10.1016/S0303-2434(02)00002-8.

Rinner, Claus, and Victoria Fast. 2015. "A Classification of User Contributions on the Participatory Geoweb." In *Advances in Spatial Data Handling and Analysis*, 35-49. Springer.

Rudolph, Sebastian. 2011. "Foundations of description logics." In *Reasoning Web. Semantic Technologies for the Web of Data*, 76-136. Springer.

Sabounchi, Nasim, Konstantinos Triantis, Sudipta Sarangi, and Shiyong Liu. 2011. Fuzzy Modeling of Linguistic Variables in a System Dynamics Context. Paper presented at the The 29th International Conference of the System Dynamics Society, Washington, DC, July 25 – 29, 2011.

Salas, JM, A Harth, B Norton, LM Vilches, AD León, J Goodwin, C Stadler, S Anand, and D Harries. 2011. "NeoGeo Vocabulary: Defining a shared RDF representation for GeoData." Review of. *Public draft, NeoGeo, May*.

Savino, S, M Rumor, and I Lissandron. 2009. *Road junction generalization in large scale geodatabases*: CRC Press/Balkema: London.

Schäfers, Michael, and Udo W. Lipeck. 2014. "SimMatching: adaptable road network matching for efficient and scalable spatial data integration." In *Proceedings of the 1st ACM SIGSPATIAL PhD Workshop*, 1-5. Dallas/Fort Worth, Texas: ACM.

Solutions, Vivid. 2016. "Java Conflation Suite." Accessed 18 May http://www.vividsolutions.com/JCS/.

Stadler, Claus, Jens Lehmann, Konrad Höffner, and Sören Auer. 2012. "Linkedgeodata: A core for a web of spatial open data." Review of. *Semantic Web* 3 (4):333-54.

Sterman, John. 2000. *Business dynamics : systems thinking and modeling for a complex world*. Boston: Irwin/McGraw-Hill.

Stigmar, Hanna. 2005. Matching route data and topographic data in a real-time environment. Paper presented at the 10th Scandinavian Research Conference on Geographical Information Science.

Touya, Guillaume. 2010. "A road network selection process based on data enrichment and structure detection." Review of. *Transactions in Gis* 14 (5):595-614.

Uitermark, Harry T, Peter JM van Oosterom, Nicolaas JI Mars, and Martien Molenaar. 1999. Ontology-based geographic data set integration. Paper presented at the Spatio-temporal database management.

Vandenbroucke, Danny, Joep Crompvoets, Glenn Vancauwenberghe, Ezra Dessers, and Jos Van Orshoven. 2009. "A Network Perspective on Spatial Data Infrastructures: Application to the Sub-national SDI of Flanders (Belgium)." Review of. *Transactions in Gis* 13 (s1):105-22.

Verburg, R. M., and J. H. Andriessen. 2006. "The assessment of communities of practice." Review of. *Knowledge and Process Management* 13 (1):13-25.

Volz, S. 2006. An iterative approach for matching multiple representations of street data. Paper presented at the ISPRS Workshop, Multiple representation and interoperability of spatial data, Hanover, Germany.

W3C Semantic Web Interest Group. "Basic RDF Geo Vocabulary: Point/lat/long/alt." http://www.w3.org/2003/01/geo/#vocabulary.

Walter, Volker, and Dieter Fritsch. 1999. "Matching spatial data sets: a statistical approach." Review of. *International Journal of Geographical Information Science* 13 (5):445-73.

Weiss, Roy, and Robert Weibel. 2014. "Road network selection for small-scale maps using an improved centrality-based algorithm." Review of. *Journal of Spatial Information Science* 2014 (9):71-99.

Wenger, E. 2000. "Communities of practice and social learning systems." Review of. *Organization* 7 (2):225-46.

Wenger, E., R. A. McDermott, and W. Snyder. 2002. *Cultivating communities of practice: A guide to managing knowledge*. Harvard Harvard Business Press.

Wiki, Licence-OSM foundation. Accessed 14, Sep, 2016. http://opendatacommons.org/licenses/odbl/.

Will, Julian. 2014. "Development of an automated matching algorithm to assess the quality of the OpenStreetMap road network: a case study in Göteborg, Sweden." Review of. *Student thesis series INES*.

Wille, Rudolf. 1992. "Concept lattices and conceptual knowledge systems." Review of. *Computers & Mathematics with Applications* 23 (6):493-515. doi: http://dx.doi.org/10.1016/0898-1221(92)90120-7.

Yang, Bisheng, Xuechen Luan, and Qingquan Li. 2010. "An adaptive method for identifying the spatial patterns in road networks." Review of. *Computers, Environment and Urban Systems* 34 (1):40-8. doi: http://dx.doi.org/10.1016/j.compenvurbsys.2009.10.002.

———. 2011. "Generating hierarchical strokes from urban street networks based on spatial pattern recognition." Review of. *International Journal of Geographical Information Science* 25 (12):2025-50. doi: 10.1080/13658816.2011.570270.

Yang, Bisheng, Xuechen Luan, and Yunfei Zhang. 2014. "A Pattern-Based Approach for Matching Nodes in Heterogeneous Urban Road Networks." Review of. *Transactions in Gis* 18 (5):718-39.

Yang, Bisheng, Yunfei Zhang, and Xuechen Luan. 2013. "A probabilistic relaxation approach for matching road networks." Review of. *International Journal of Geographical Information Science* 27 (2):319-38.

Yang, Lin, Zejun Zuo, Run Wang, Yaqin Ye, and Maosheng Hu. 2014. "Matching Road Network Combining Hierarchical Strokes and Probabilis-tic Relaxation Method." Review of. *Open Automation and Control Systems Journal* 6:268-76.

Zadeh, L. A. 1973. "Outline of a New Approach to Analysis of Complex Systems and Decision Processes." Review of. *Ieee Transactions on Systems Man and Cybernetics* Smc3 (1):28-44. doi: Doi 10.1109/Tsmc.1973.5408575.

Zhang, Meng, and Liqiu Meng. 2008. "Delimited stroke oriented algorithm-working principle and implementation for the matching of road networks." Review of. *Geographic Information Sciences* 14 (1):44-53.

Zhang, Meng, Liqiu Meng, and Joachim Bobrich. 2010. "A road-network matching approach guided by 'structure'." Review of. *Annals of GIS* 16 (3):165-76. doi: 10.1080/19475683.2010.513154.

DEPARTMENT OF PHYSICAL GEOGRAPHY
AND ECOSYSTEM SCIENCES
FACULTY OF SCIENCE
LUND UNIVERSITY