Chapter 1 Introduction

1.1 An analysis of a data grid approach for spatial data infrastructures

Grid computing started in the 1990s as a future generation computing paradigm for high performance computing. The initial goals were to extend processing and data storage capacities from individual expensive machines to clusters of inexpensive commodity machines, mainly for use in the scientific domain. The vision was to create a 'grid' of networked computers into which anyone could tap for processing and data storage capacity, analogous to a power grid into which we tap for electrical power (Foster and Kesselman 1999). Some ideas originating from grid research have permeated into all areas of distributed computing, changing the way in which distributed systems are designed, developed and implemented by addressing the needs for flexible, secure, coordinated resources sharing among members of a virtual organization comprising individuals, institutions and resources from different administrative domains (Foster *et al.* 2001, Talia 2002, Ripeanu *et al.* 2008).

Most grids have a *service-oriented architecture* and there is close cooperation with the world of *web services* (Foster 2003, Baker *et al.* 2005, Cohen *et al.* 2008), which are software systems that support interoperable machine-to-machine interaction over a network (Haas and Brown 2004). Grid and web service technologies complement and influence each other, and since both are fairly young it is entirely possible that in future they will become fully compatible and the distinction between the two will fade (Plaszczak and Wellner 2006) so that at some point in future they might be known under a single name. Grid computing research has also been the breeding ground for new technologies known under different names, such as, cloud computing, the latest catchphrase in industry, which shares the same original vision of grid computing articulated in the 1990s by Foster, Kesselman and others, but with significant differences (Weiss 2007, Delic and Walker 2008).

Over the past few years 'geobrowsers', such as Google Earth, NASA World Wind and Virtual Earth along with in-vehicle navigation, handheld GPS devices and maps on mobile phones, have made interactive maps and geographic information an everyday experience. Behind these maps lies a wealth of spatial data that is often collated from a vast amount of different sources. Consolidating spatial data from distributed heterogeneous sources into a single centralized dataset that can be published online is a time consuming effort, requiring, among others, a considerable coordination effort, as well as syntactic and semantic data harmonization. A *spatial data infrastructure (SDI)* aims to make spatial data usable by people, and the technologies, systems (hardware and software), standards, policies, agreements, human and economic resources, institutions, and organizational aspects have to be carefully orchestrated to make this possible (Jacoby 2002, Crompvoets *et al.* 2004, Georgiadou *et al.* 2005, De Man 2006, Rajabifard *et al.* 2006, Masser *et al.* 2007). SDI research provides insights into understanding and improving the consolidation of heterogeneous distributed databases and making these available to as wide an audience as possible (Williamson *et al.* 2006, Masser *et al.* 2007, Rajabifard 2008).

This dissertation spans two disciplines, namely Computer Science (CS) and Geographic Information Science (GISc). The data grid approach (CS) as the enabling technology for sharing geographic information, such as address data, in an SDI (GISc) is presented and analyzed. This first chapter introduces the reader to address data in an SDI (GISc) and to data grids (CS), and then presents two scenarios (developed by the author) that illustrate how data grids could in future enable the sharing of address data in an SDI. Subsequently, the research presented in this dissertation is related to current research agendas in the two disciplines. The chapter is concluded with an overview of the contributions from the work described in this dissertation to scientific research, and a guide for the reader to the remaining chapters of the dissertation. Figure 1 illustrates how the chapters in this dissertation relate to the two disciplines.

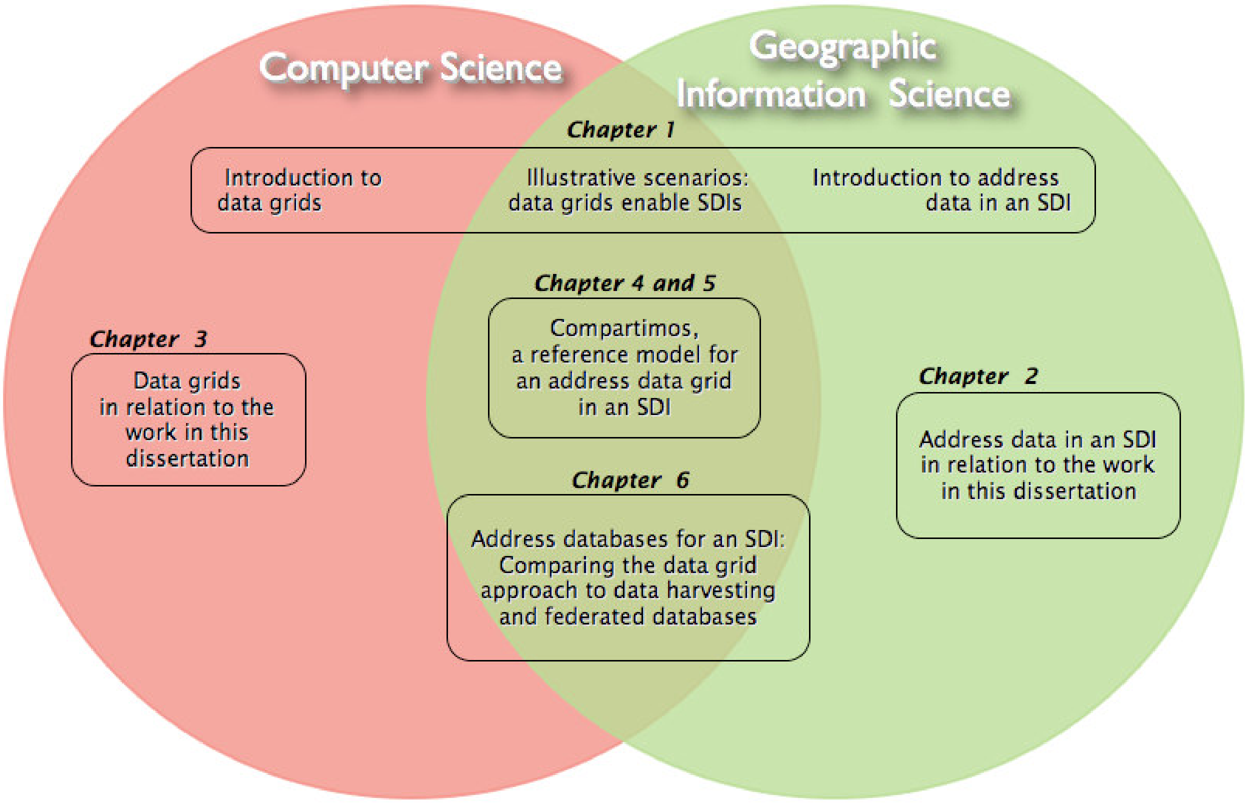
Computer Science
Geographic Information Science
Chapter 1
Introduction to data grids
Illustrative scenarios: data grids enable SDIs
Introduction to address data in an SDI
Chapter 4 and 5
Compartimos, a reference model for an address data grid in an SDI
Chapter 3
Data grids in relation to the work in this dissertation
Chapter 2
Address data in an SDI in relation to the work in this dissertation
Chapter 6
Address databases for an SDI: Comparing the data grid approach to data harvesting and federated databases

Figure 1. The chapters in this dissertation in relation to CS and GISc

1.2 Address data in an SDI

The original purpose of numbered street addresses was to enable the correct and unambiguous delivery of letters and parcels, i.e. postal services. This purpose is reflected in the definitions for an address found in many English dictionaries, of which two are shown in Table 1 below.

Table 1. Address definitions

Oxford English Dictionary	the direction or superscription of a letter, etc.; the name of the person and place to which it is addressed or directed; the name of the place to which any *one's letters are directed*. (Oxford University Press 2007b)
Cambridge Advanced Learner's Dictionary	the number of the house and name of the road and town where a person lives or works and *where letters can be sent*. (Cambridge University Press 2007)

However, in this dissertation an address is regarded in its broader sense as the description of a location not only for postal delivery, but for *all* kinds of service delivery, ranging from "physical" services such as utility services (water, electricity, sewerage, etc.), billing, courier, goods delivery, and emergency dispatch; to more "abstract" services such as opening financial accounts, credit application, tax collection, and land and property registration (Coetzee and Cooper 2007b). Any information about the recipient of the service is delivery, whether a person or an organization, is not included in the address. Table 2 lists a few sample addresses, including some that are not valid for postal delivery.

Table 2. Sample addresses

South Africa	Corner Kings and Richmond Roads Mowbray Cape Town	Germany	Waldparkstrasse 67c Hamburg
Japan	14F Sphere Tower Tennoze 2-2-8 Higashishinagawa Shinagawaku Tokyo 140 0002	Spain	Calle Agazado, 23 Molino de la Hoz Las Rosas ES-28230 Madrid

Address data refers to a collection of addresses, and *reference data* to data according to which other information can be referenced unambiguously. Thus, *address reference data* is a collection of addresses according to which other information can be referenced unambiguously. As an example, in a corporate database each customer could be referenced to an address, or in a disaster management situation specific incidents could be referenced to an address. This implies that the address is an independent entity or object to which the other information is linked. This dissertation is about address reference data but for simplicity reasons the term 'address data' is used when referring to 'address reference data'.

Due to their service, infrastructure and land administration responsibilities, it is commonly found that a local authority establishes and maintains address data for its area of jurisdiction (Levoleger and Corbin 2005, Williamson *et al.* 2005, Coetzee *et al.* 2008b). When address data is required for an area that extends across these jurisdictional boundaries, the data has to be collated from the various local sources. For this reason, address data is part of a country's SDI, the infrastructure that is required to make spatial data from various sources useful and available to as wide an audience as possible. In this dissertation a novel approach for dynamically consolidating and sharing address data from multiple sources is presented.

1.3 Data grids

Grid computing started in the late 1990s as a distributed infrastructure for specific Grand Challenge applications, the main purpose being high performance computing. Since then it has expanded to address the general need for flexible, secure, coordinated resource sharing among collections of individuals, institutions and resources (Foster *et al.* 2001). Apart from computing resources, a grid can also share data or storage resources, or provide access to sensors and/or specialist equipment such as particle accelerators used in physics experiments. There is an abundance of definitions for a grid, and one that is found very often, is Foster's (2002) three point check list, stating that a grid is a system that:

1. coordinates resources that are not subject to centralized control;
2. delivers non-trivial qualities of service; and
3. uses standard, open, general-purpose protocols and interfaces.

That is, the individual resources that are shared on a grid live in different control domains, for example, different institutions or different administrative divisions of the same institution; the constituent resources of the grid are coordinated to deliver a service that is significantly greater than the sum of its parts and, all of this is achieved through standard, open, general-purpose protocols and interfaces.

A *data grid* is a special kind of grid in which mainly data resources are shared. That is,

1. the individual datasets that are shared on the grid live in different control domains and consist either of files or of databases created and maintained within a database management system (DBMS), or of both;
2. these constituent datasets are coordinated to deliver a virtual dataset (service) that is significantly greater than the sum of its parts and,
3. all of this is achieved through standard, open, general-purpose protocols and interfaces.

A data grid is the platform for data sharing in a virtual organization consisting of individuals and/or institutions that work together for collaborative problem solving or other purposes. Data grids are implemented to enable *data federation*, i.e. the logical integration of multiple data services or data resources so that they can be accessed as if they were a single service (OGF 2007c); and/or data grids are implemented in *data-intensive environments* to enable efficient access to, and the movement and management of, large quantities of data in a distributed environment (Chervenak *et al.* 2000, Venugopal *et al.* 2006). The work described in this dissertation gravitates towards data federation, but borrows from replication and data transfer, as they would be used in a data-intensive data grid. In this dissertation the data grid approach is presented as a novel way to enable the sharing of address data in an SDI environment.

1.4 Enabling spatial data infrastructures with data grids

A local authority usually maintains address data for its area of jurisdiction. When address data is required for a larger area, at first glance, the obvious solution is to aggregate the address data from the individual local databases into a single centralized database. This approach is followed in countries such as Australia, Ireland, the United Kingdom and Denmark (Paull 2003, Lind and Nicholson 2004, Fahey and Finch 2006). The percentage of participating local authorities varies considerably in these countries. A number of studies have shown that the sharing and collation of local spatial data, including address data, is not yet common. These studies have found that:

- the involvement in SDIs of an increasing number of participants from all levels of government as well as the private sector has resulted in *generally uncoordinated activity* (Rajabifard *et al.* 2006, Williamson *et al.* 2006);
- a *federated approach* to data sharing is more sustainable than a centralized approach (Harvey and Tulloch 2006, Carrera and Ferreira 2007);
- the bottom-up (involving all levels of government as well as the private sector) approach to an SDI results in a *large diversity and heterogeneity of stakeholders and their resources at disposal* (Rajabifard *et al.* 2006, Williamson *et al.* 2006); and
- there is an *increased demand for spatial data* (and thus spatial data sharing) due to the use of state-of-the-art consumer technology such as Internet mapping and routing sites, GPS devices, and in-vehicle navigation (Williamson *et al.* 2006, de Man 2007, Craglia *et al.* 2008); and
- *data sharing among SDI participants on an unprecedented scale* is needed for SDIs to become fully operational and effective in practice (Rajabifard *et al.* 2006, Masser *et al.* 2007).

Web services support interoperable machine-to-machine interaction over a network (Haas and Brown 2004), and are therefore ideal to enable decentralized access to distributed data on heterogeneous platforms. Grid technologies evolved from custom solutions and the early versions of the Globus Toolkit in the 1990s to the Open Grid Services Architecture (OGSA) of the 21st century which aligns Grid computing with service-oriented architectures and Web services, and provides a reference model within which one can define a wide range of interoperable, portable services (Foster and Kesselman 2004). OGSA includes the description of Web services for data management with the functionality for storage, movement, access, replication, caching and federation of files and databases (OGF 2007c). These Web services for data management comprise the essential capabilities that are required to make individual heterogeneous datasets appear as a single virtual dataset, i.e. these services enable data federation in a data grid. This dissertation analyzes the use of these data grid services for solving the problem of sharing spatial data, such as address data, in an SDI environment.

Venugopal *et al.* (2006) described a number of characteristics that are unique to data grids, such as geographically distributed and heterogeneous resources under different administrative domains, and a large number of users sharing these resources and wanting to collaborate with each other. These data grid characteristics are similar to the data sharing challenges facing SDIs, mentioned in numerous SDI research papers (Georgiadou *et al.* 2005, McDougall *et al.* 2005, Tuladhar *et al.* 2005, Williamson *et al.* 2005, Rajabifard *et al.* 2006, Masser *et al.* 2007, Craglia *et al.* 2008). This similarity shows that there is a pre-existing link between the problem of data sharing in an SDI and the solution that a data grid provides.

Adapting Foster's (2002) definition of a grid, an *address data grid* would be the following:

1. the individual address datasets that are shared live in different control domains *without centralized control*, i.e. at the different local authorities and other institutions;
2. these constituent address datasets are coordinated to deliver access to a virtual address dataset (service) that is significantly greater than the sum of its parts (across jurisdictions), a *non-trivial quality of service*; and
3. all of this is achieved through *standard, open, general-purpose protocols and interfaces.*

In this dissertation the novel approach of a data grid is explored as the enabling platform for an SDI so that address data can be consolidated and shared at national or international level. The focus is on issues that are relevant to address data in an SDI environment, as opposed to other data under different circumstances. These issues include *the nature of address data production and maintenance*, address-related *services* that justify a data grid and *standards and protocols* that are required to seamlessly integrate address data from multiple sources.

In the following sections 1.4.1 and 1.4.2 two illustrative scenarios (developed by the author) for an address data grid in an SDI illustrate how a data grid could in future enable the consolidation of address data from multiple sources, and also provide services that are beyond the capacity of an individual organization. Similar to theses scenarios, Plaszczak and Wellner (2006) use examples to illustrate what Grid technology in general (as opposed to specifically for address data in an SDI) can deliver in the near future.

1.4.1 Scenario 1: A deadly storm hits the border between two countries

A deadly storm with high winds and heavy rains hits an area that is on the border of two countries. An emergency response centre (ERC) immediately starts operating and starts receiving reports of damage sites and people in distress from the various sources, including the public. In order to be prepared, the ERC maintains a computing infrastructure that is adequate for the worst imaginable disaster. The ERCs demand for computing infrastructure peaks during the emergency response phase of a disaster and in relation, in between disasters, the demand for computing infrastructure is extremely low. During a disaster, the ERC maps the incidents and provide maps with locations of distress and damage to the rescue and clean-up teams. In urban areas the damage sites and distress locations are mostly referenced by address. In rural areas distress locations are less frequently reported as addresses, but more often as descriptions of locations.

To map the location of damage or distress reports, the address on an incoming report is matched to an address in an address dataset that includes geo-spatial coordinates, a process known as geocoding. The ERC is in possession of software that automates the geocoding but it requires the address data to be in a single database, structured according to a specific data model. The address data is also used as backdrop for any maps that are sent to the rescue and clean-up teams. Address data has to be collected from the 50 odd individual cities and towns that have been affected by the storm. In the one country an aggregated address dataset exists but an updated version of the data is released only every six months. In the other country the data is available at individual local authorities only, where each dataset is based on a different data model that includes city-specific semantics. The datasets are prepared and released in the proprietary data format of software from different vendors. For most cities, the address data can be viewed on an Internet mapping site but a city's complete dataset cannot be downloaded. As a rule, a city's complete dataset is available on disk for emergency response and disaster management, but some cities require a signature for receipt of the data to ensure that their address data is used for those purposes only.

Without the option of a data grid, the ERC has to collect the data from the individual cities and towns, where possible electronically (e.g. downloaded from an ftp site), otherwise physically by sending a messenger to collect a disk, and then proceed in one of three ways.

The first option comprises of converting the address data from each city into the data model required for the geocoding software and loading this converted data into a single database. This is a time consuming process and any anomalies have to be manually resolved (increasing the turnaround time) or are rejected (reducing the size and coverage of the dataset). By the time this process is completed, everybody might have forgotten about the disaster.

A second option is to set-up the geocoding tool to work for each of the 50 different data formats from the individual cities and towns, i.e. 50 different configurations of the geocoding software. This slows down the geocoding process, since incoming addresses have to be assigned to a city or town before geocoding can proceed.

A third option is to not use an automatic geocoding tool, but rather to add individual datasets to one large map on which geocoding is done manually by humans interpreting and finding the address on the map. These manual searches can take up to a few minutes per address in a metropolitan area. If there are many distress reports coming in simultaneously, these few minutes could mean the difference between life and death.

Projecting this scenario into a future world where an address data grid is a reality, the following is possible.

Applications, processing cycles and datasets are abstracted as resources in a Grid world. Each resource can be accessed remotely according to its individual policy. Thus each city can securely grant rights to the emergency response centre for access to its address dataset. This eliminates the need to download data or physically collect data, while at the same time protecting the privacy and integrity of the city's data. An address data grid also requires standardization in terms of address data exchange. Even though each city maintains its data according to its own data model, it publishes and makes available a Grid-enabled Web service, or a Grid service, that provides access to its address data according to an agreed upon address data exchange standard and protocol. The geocoding software makes use of these Grid services to seamlessly work with the data from any city. In this way the ERC is guaranteed to display the latest address data on the map, which is important if the disaster strikes newly developed areas. Figure 2 shows how the different components interact in this scenario.

The cities can further configure their spare processing cycles as Grid resources that can be used by the ERC during a disaster, alleviating the centre from the burden of maintaining a computing infrastructure that is only used occasionally. Alternatively, the ERC maintains a computing infrastructure that is adequate for the worst imaginable disaster and rents out the processing cycles as Grid resources in between disasters, thereby providing a better justification for the initial capital investment.

Further, if the geocoding software is Grid-enabled, it can execute in parallel on the grid processing resources of the different cities. Thus once the city for an incoming address is known, the address can be sent to the grid processing resource of that city where the rest of the geocoding is performed. When an address needs to be matched, the process of matching the city is fast in comparison to matching the combination of suburb, street and street number, and the latter part of the match that takes up more time is then processed in parallel. Such a strategy increases address throughput because addresses are now matched in parallel against a smaller reference dataset (single city) and the address is matched close to the list of potentially matching addresses.

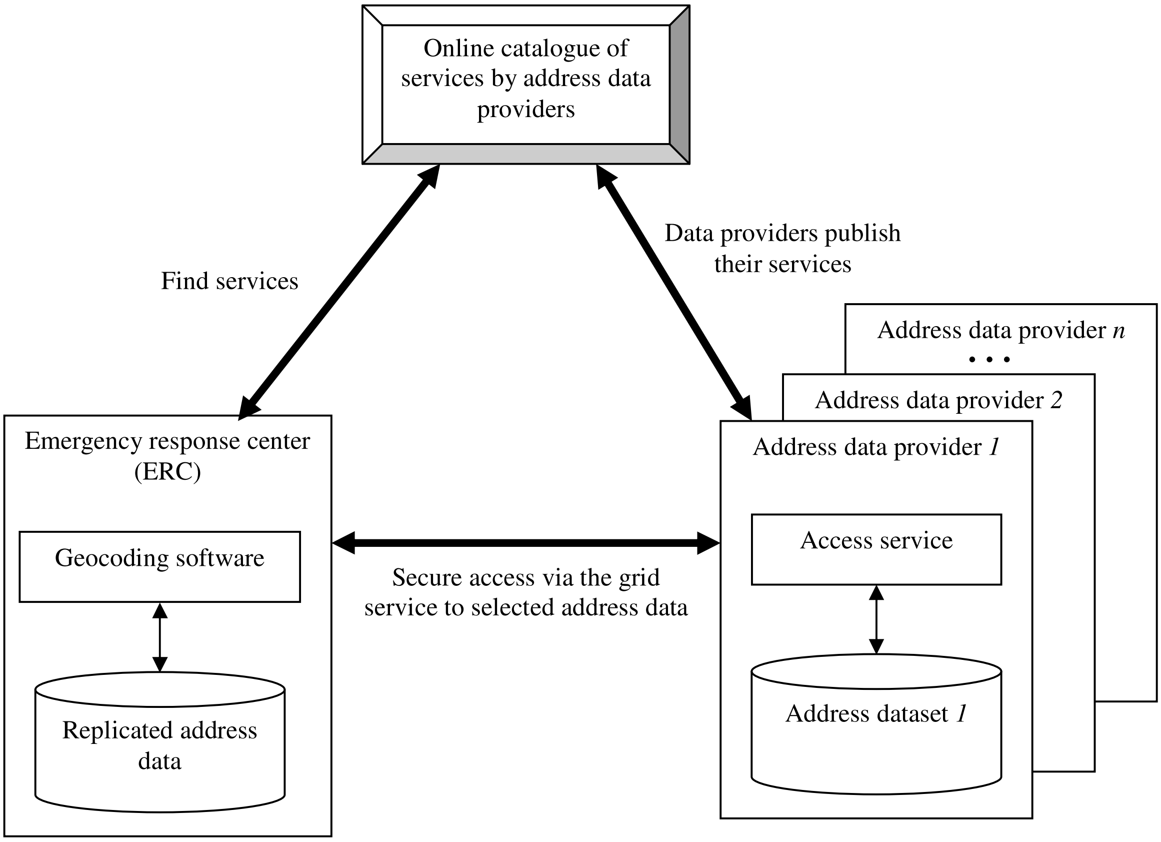
Online catalogue of services by address data providers
Find services
Data providers publish their services
Address data provider *n*
• • •
Address data provider *2*
Address data provider *1*
Access service
Address dataset *1*
Emergency response center (ERC)
Geocoding software
Replicated address data
Secure access via the grid service to selected address data

Figure 2. Scenario 1 – Mapping the locations of damage and distress reports

Naturally, when disaster strikes, it does not affect a single address location but an area comprising numerous address locations. Thus an alternative strategy would be the following: when a geocoding request is sent to a city's address data, the address reference for that suburb and its neighborhood is immediately replicated at the emergency response centre. Subsequent geocoding requests from that area are then processed locally (and therefore faster) at the ERC.

With the data grid the ERC gets access to the latest up-to-date consolidated address data for automatic geocoding, eliminating manual intervention; secondly, it either saves on computing infrastructure or gets a better return on investment on the initial capital investment; and lastly the cities can control that their address data is accessed securely for the purposes of emergency response only.

1.4.2 Scenario 2: Property valuation

An airline rewards company, AirMiles, wants to introduce an AirMiles credit card to its estimated ten million international customers. They have contacted FinBank as the provider of the credit card. FinBank are interested, but they want to evaluate the customer base before finalizing the terms and conditions and signing an agreement. This evaluation includes a valuation of the property at each AirMiles customer's residential address. The property valuation comprises geocoding the customer's address and comparing it to other datasets such as credit rating per suburb, soil features of the area, and proximity to the public transport network. Neither AirMiles nor FinBank are experts in these areas and have contracted ConsultCo to do the property valuation.

The AirMiles customer base spans more than one country and therefore the geocoding has to be done against address data collected from different countries, including local authorities within these countries. In some countries this data is available for free, in others the data has to be purchased. Since customers are randomly spread across the country, it is not known which parts of the country are needed for the geocoding, and therefore the dataset for the whole country has to be purchased, where applicable, at a steep fee. The AirMiles customer database in itself is a valuable asset that has to be protected and it includes personal information about customers that requires protection for privacy reasons. AirMiles would prefer employees from ConsultCo doing the valuation on-site at the AirMiles offices where stringent security measures are in place. This implies that ConsultCo have to fly in experts from their different offices, adding to the traveling costs. Finally, the licensing of the sophisticated geocoding software package that ConsultCo uses, does not allow ConsultCo to install the geocoding software on AirMiles machines. The property valuation is quite simple, but the geocoding depends on an address dataset spanning more than one country without which the rest of the valuation cannot continue.

Again, if this scenario is projected into a future world where an address data grid is a reality, the valuation process could be simplified as follows.

AirMiles configure their customer database as a Grid resource for which they set a strict policy that allows ConsultCo access to a customer's address for purposes of geocoding only, and one or two attributes of a customer into which they can write geocoding and valuation information. ConsultCo queries an online directory of address data providers who have set-up their address datasets as Grid resources and provide access to their data through standardized Grid services. A specific address data provider could supply data for an area ranging from a local authority's jurisdiction to a province or state, a country or even an international region. The online directory includes pricing and quality of service information so that ConsultCo can pick the best offer available. The Grid services are standardized to eliminate differences resulting from data stored in underlying DBMSs from different vendors. The Grid services are further standardized to exchange address data in a standard format that the ConsultCo geocoding software understands.

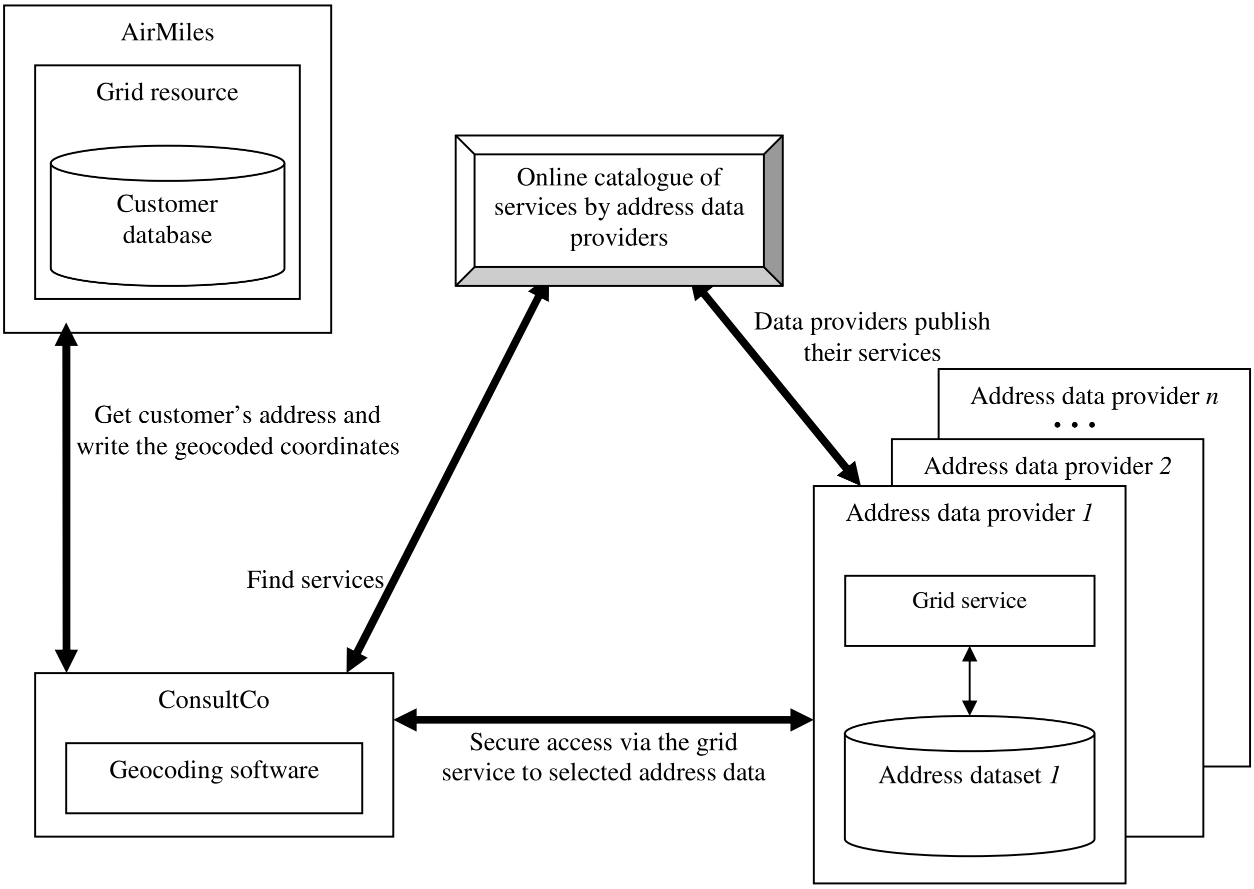
AirMiles
Grid resource
Customer database
Online catalogue of services by address data providers
Data providers publish their services
Address data provider *n*
• • •
Address data provider *2*
Address data provider *1*
Grid service
Address dataset *1*
Get customer's address and write the geocoded coordinates
Find services
ConsultCo
Geocoding software
Secure access via the grid service to selected address data

Figure 3. Scenario 2 – Geocoding a customer database

ConsultCo now executes their geocoding software from machines at their offices, which reads the customer address from the server at AirMiles offices, matches it to the address data providers from the relevant country, and writes the resulting coordinate into the geocoding attribute in the AirMiles customer database. When it is time for the property valuation, ConsultCo access the customer geocoding attribute (coordinates) and compare it to the other datasets of credit ratings, soil features and the public transport network (which in turn could each come from a different grid resource). The resulting valuation information is written into the valuation attribute on the AirMiles customer database. Refer to Figure 3 for this scenario.

There is no need for ConsultCo employees to be on-site at the AirMiles offices and the logistics are simplified as the employees can continue with the work from the desktops in their respective offices. ConsultCo does not have to purchase the address data for the whole country, nor does it have to consolidate the data from multiple sources, rather it only uses and/or pays for the specific data that is required to geocode the addresses. Thus, the data grid has simplified the logistics and therefore the costs of the project, and more importantly the costs and network traffic for the address data have been significantly reduced since ConsultCo accesses relevant address data only.

1.5 Computer Science and Geographic Information Science in this dissertation

This dissertation spans two disciplines: Computer Science (CS) and Geographic Information Science (GISc), and this section relates the work described in this dissertation to current research initiatives in these two disciplines.

1.5.1 Computer Science

The work described in this dissertation relates to current research on *data grids* and *distributed computing* in Computer Science. Based on the literature review that was done as part of this dissertation, the author has identified four categories of current grid research in Computer Science publications:

1. The *philosophy* behind the idea of the Grid, the *fundamental principles* of the Grid idea and the *motivation* for Grid technology. Foster's (2002) *What is the Grid? A three point checklist*, Part I of the second Grid book edited by Foster and Kesselman (2004), and the first part of the book by Berman *et al.* (2003) are examples.

2. Grid *concepts, architecture and technologies* describe the inner workings of a grid, as well as tools and technologies that are used in a grid environment. *The Anatomy of the Grid* (Foster *et al.* 2001), *The Physiology of the Grid* (Foster *et al.* 2002), the publications by the Open Grid Forum (OGF) regarding the Open Grid Services Architecture (2006) and Part II, IV, V and VI of the second Grid book by Foster and Kesselman (2004) fall into this category.

3. Grid *applications*. This research includes reports and lessons learnt from Grid implementations in various disciplines and domains. A vast amount of reports and papers have been published and the following are examples: Bernholdt *et al.* (2005) on climate modeling research, Gomez-Iglesias *et al.* (2008) on physics experiments, Volckaert *et al.* (2008) on media production and distribution and Chu *et al.* (2008) on a Grid computing platform for human and animal kidney research. Furthermore, Chapters 3 to 6 of the first Grid book by Foster and Kesselman (1999) are dedicated to application

domains, as are Part II of the second Grid book by Foster and Kesselman (2004), Part D of the book edited by Berman *et al.* (2003). From these reports one can abstract general features of different types of applications well suited for the Grid environment.

4. *Low-level infrastructure topics related to the Grid* such as data replication, resource management, workflow management and job scheduling. This kind of research explores specialized topics in a Grid environment, for example, comparing different strategies and algorithms to schedule data replication in a grid environment. Once again publishing activity in this area is high, and the following are examples: Li and Buyya on grid scheduling strategies (2007), De Rose *et al.* (2008) allocation strategies, Rabl *et al.* (2008) on dynamic allocation in a self-clustering database and Bruin *et al.* (2008) on job submission.

Research in this dissertation falls into the third category, namely Grid applications. The application environment is described in Chapter 2 where the characteristics of the *SDI* environment in which address data is produced, maintained and shared are described. Chapter 3 interprets current data grid research in relation to the work described in this dissertation and concludes with a discussion of related research, confirming that the work described in this dissertation is innovative and new, but also extremely relevant at the current point in time. Compartimos, a reference model for an address data grid, is presented in Chapters 4 and 5. Compartimos is based on the OGSA data architecture and aims to provide a solution for a problem on the application layer, i.e. data sharing in an SDI. In Chapter 6 the data grid approach to the application area of national address databases in an SDI is compared to more traditional approaches. Aspects of Computer Science that require further research, forthcoming from the work in this dissertation, are discussed in Chapter 7.

1.5.2 Geographic Information Science

Information integration, distributed computing and *SDIs* have been identified as priority areas for research in Geographic Information Science (Goodchild *et al.* 2005, Onsrud *et al.* 2005, Craglia *et al.* 2008). In 'A research agenda for geographic information science' Goodchild *et al.* (2005) recognize distributed and mobile computing as a significant area of Geographic Information Science research. They note that there is widespread interest among members of the Geographic Information Science community in the support for true distributed databases, such that a user sees a single database, but tables or even parts of tables are resident at different server sites. A detailed map for a country is one such example, where the street network and address data for each town or city is maintained and made available by the individual local authorities, but displayed on a single national map.

In the position paper by Craglia *et al.* (2008), a group of international geographic and environmental scientists from government, industry and academia argue that the vision of Digital Earth put forward by US Vice-President Al Gore 10 years ago needs to be re-evaluated in the light of the many developments in the fields of information technology, data infrastructures, and earth observation that have taken place since. The position paper focuses the vision on the next generation Digital Earth and identifies priority research areas to support this vision. These priority areas include information integration (multi-source and heterogeneous, multi-disciplinary, multi-temporal, multi-resolution, multi-lingual and multi-media) and computational infrastructures to implement this vision (architecture, data structures, indexing and interfaces). Both these priority areas are addressed in this dissertation.

The list of components of an SDI varies in literature (Crompvoets *et al.* 2004, Georgiadou *et al.* 2005, De Man 2006, Williamson *et al.* 2006, Masser *et al.* 2007, Rajabifard 2008), but generally includes the following technical and non-technical aspects:

- *Technical aspects*: Technologies, systems (hardware and software) and standards.
- *Non-technical aspects*: Policies, agreements, human and economic resources, institutions and organizational aspects.

Refer also to Figure 5 on p25, which illustrates how these aspects together make spatial data available to people. The work described in this dissertation relates to the technical aspects of an SDI. The non-technical aspects are obviously invaluable for a successful SDI, but they are out of scope for this dissertation.

The work described in this dissertation sheds light on the use of a data grid for the kind of *distributed* database that is described by Goodchild *et al.* In Chapter 2 the characteristics of the *SDI* environment in which address data is produced, maintained and shared are described. Compartimos (Spanish for ‘we share’), a reference model for an address data grid, is presented in Chapters 4 and 5. Compartimos is based on the OGSA data architecture and is an abstract representation of the essential components required to enable data sharing in an SDI with data grids. The novel address data model that is described in Chapter 4 illustrates the importance of application domain specific standards for data *integration*. The benefits of the novel data grid approach for address databases for national SDI are highlighted in Chapter 6 and new research questions in Geographic Information Science, initiated from the work in this dissertation, are discussed in Chapter 7.

1.6 Contributions to scientific research from this dissertation

The main results and contributions from this dissertation towards the two disciplines of Computer Science and Geographic Information Science are discussed in the following paragraphs.

Vision of an address data grid in an SDI. The first contribution from this dissertation lies in the two scenarios described earlier in this chapter. These scenarios illustrate for the first time how data grids can be applied to enable the sharing of address data in an SDI and thus illustrate the vision of an address data grid in an SDI. The scenarios enhance the mutual understanding of the grid computing and geographic information domains by describing how data grids can solve two very real problems in the geographic information domain. Further to this mutual understanding, in section 1.5, the work described in this dissertation is related to current research initiatives in Computer Science and Geographic Information Science to confirm that an address data grid in an SDI fits into the research agendas of both disciplines. The work in this dissertation is part of Grid application research and falls under the identified GISc research priority areas of information integration, distributed computing and SDIs.

Understanding what an address is. The definition of an address in the broader sense (i.e. not only for postal delivery), the notion of an address as a reference, instead of being an attribute, and the definition of an addressing system and the comparison to spatial reference systems, as presented in Chapter 2, enhance the understanding of what an address is. This provides important groundwork for the novel Compartimos address data model. The environment in which address data is produced and maintained in an SDI is discussed in Chapter 2, as well as in the enterprise viewpoint of Chapter 4, and more specifically for South Africa in Chapter 6. These discussions confirm that address data should be seen in the context of an SDI.

Similarities between data grids and address data in an SDI. In relation to Foster's (2002) three-point checklist for a grid system, similarities between SDI address data sharing and data grids have been identified for the first time. These similarities are discussed in Chapters 3 and 6 and further enhance the mutual understanding of the grid computing and geographic information domains. A novel evaluation framework for national address databases in an SDI was developed and is presented in Chapter 6. The data grid, as well as other models, was evaluated against this framework, and this evaluation enhances the understanding of the benefits that the data grid approach brings to national address databases in an SDI, as described in Chapter 6.

Reference model for an address data grid in an SDI. The Compartimos reference model, presented in Chapters 4 and 5, is a first attempt at identifying the components with their capabilities and relationships that are required to realize an address data grid in an SDI. Compartimos advances the understanding of the requirements for, and the use of, the data grid approach in a specific application domain, namely address data in an SDI. This is both a novel application for data grids (refer to section 3.6), as well as a novel technology in SDI environments (refer to section 2.6) and thus improves the understanding of the requirements and issues related to applying Grid technology in the geographic information domain. Also adding to this understanding is the comparison between examples of existing data grid implementations and the requirements for an address data grid in an SDI in Chapter 3 and 4, as well as a description in Chapter 4 of what a virtual organization (VO) in an address data grid in an SDI would be. Also in Chapter 4, the Compartimos objects are assigned to one of the layers in the Grid architecture, illustrating at which level of abstraction application domain-specific integration is required.

The novel address data model that is presented in Chapter 4 shows the importance of application domain-specific standards for data integration and is an example of what an international standard for address data exchange could look like. The model shows that it is possible to design a data model for sharing and exchange, despite diverse addressing systems and that it does not impact on, or interfere with, local laws regarding address allocation. Compartimos further confirms the need for standardization of domain specific geographic information, such as address data, and their associated services in order to integrate data from distributed heterogeneous sources.

The technology choices for a Compartimos implementation, described in Chapter 5, analyze the usability of existing technologies in Compartimos, identifying how these technologies can be applied to Grid-enable address data sharing in an SDI.

Recommendations and new questions. The discussion of Compartimos in Chapter 5 proposes expansions to the Compartimos reference model. All in all, the contributions of this research have led to new questions, such as the viability of cloud computing for data sharing in an SDI and the involvement of the community in maintaining address data, as described in Chapter 7. These questions have to be addressed through further research.

1.7 Guide to the remaining chapters of this dissertation

The remaining chapters of this dissertation are described below. Refer also to Figure 1 on p2 for a graphic illustration of the chapters in this dissertation in relation to Computer Science and Geographic Information Science.

Chapter 2 – This chapter provides information about address data in the context of SDIs to show that the data grid approach is a novel way of addressing the problem of address data sharing in an SDI. Definitions for the terms ‘address’, ‘addressing system’ and ‘address reference data’ are provided to enhance the understanding of what an address is. The chapter includes a discussion on SDIs, why it is important to consider address data in the context of an SDI and of the similarities between SDI address data sharing and SDIs. An overview of technologies and standards currently used in SDIs provides input to the technology choices discussed in Chapter 5. The chapter is concluded with a discussion of work related to the research in this dissertation to point out similarities and to highlight the novelty and uniqueness of this research in the GISc discipline.

Chapter 3 – In this chapter more information about grid computing and data grids is provided with the goal of showing that the data grid approach as enabler for SDI data sharing is both innovative and new, and also extremely relevant at the current point in time. A few existing data grid implementations were chosen to highlight similarities and differences between those applications and the work described in this dissertation. The chapter is concluded with an overview and discussion of current research work that is related to the work described in this dissertation, confirming that the work is innovative and new, but also extremely relevant at the current point in time.

Chapter 4 – In this chapter Compartimos, a reference model for an address data grid in an SDI environment, is presented in terms of the first four of the five viewpoints of the ISO Reference Model for Open Distributed Processing (RM-ODP), i.e. the enterprise, information, computational and engineering viewpoints. Compartimos is an abstract representation of the essential components and their relationships in an address data grid in an SDI environment. The chapter is concluded with a discussion of Compartimos in relation to the OGSA Data Architecture.

Chapter 5 – This chapter comprises the fifth RM-ODP viewpoint of Compartimos, the technology viewpoint, which discusses technology choices for the implementation of Compartimos. This discussion contributes towards understanding what technologies are available to make an address data grid in an SDI a reality. A proof of concept implementation of Compartimos is presented and in conclusion and Compartimos is evaluated against the novel evaluation framework for national address databases in an SDI, which is presented in Chapter 6. In conclusion, results and recommendations for future work are expansion.

Chapter 6 – This chapter comprises a paper published by the International Journal of GIS (Coetzee and Bishop 2008). The objectives and contributions of this chapter are to 1) sketch the status of spatial address data within the context of an SDI in a country like South Africa; 2) present a novel evaluation framework for national address databases; 3) describe potential information federation models for national address databases; and 4) evaluate these models according to the novel evaluation framework.

Chapter 7 – The final chapter provides a retrospective look on the work presented in this dissertation, reconfirming the contributions to scientific research from this dissertation, and finally providing recommendations for future research in this line of work.

Chapter 2 Address data in an SDI

2.1 Introduction

In the first chapter the reader was introduced to address data in an SDI. In this second chapter more information about address data in the context of SDIs is provided in order to show that the data grid approach is a novel way of addressing the problem of address data sharing in an SDI. In reference to Figure 1, this chapter relates mostly to the Geographic Information Science discipline and provides an interpretation of current GISc research in relation to the work described in this dissertation.

The chapter commences with some theory on address data in section 2.2 to clarify the broader use of the term 'address', the term 'addressing system', as well as the term 'address data' for address reference data in this dissertation. Clarification of this terminology contributes to the understanding of what an address is and provides important groundwork for the Compartimos address data model that is presented in Chapter 4. The overview of current challenges in the production, maintenance and distribution of address data in a number of countries provides a picture of the environment and challenges that have to be addressed by an address data grid in an SDI. Next in section 2.3 is a discussion of the origins, current reality, and potential future of SDIs, based on a review of literature. Compartimos contributes towards the currently emerging third generation SDIs and the future beyond. Section 2.4 explains why it is important to consider address data in the context of SDIs and why there are similarities between SDI address data sharing and data grids. Section 2.5 provides an interpretation of technologies and standards, including address standards, currently in use by actual SDIs, in preparation for the technology choices for Compartimos that are described in Chapter 5. The chapter concludes with section 2.6, a discussion of work related to the research in this dissertation to point out similarities and to highlight the novelty and uniqueness of this research in the GISc discipline.

2.2 Address data

2.2.1 Theory

The original purpose of a numbered street *address* was to enable the correct and unambiguous delivery of letters and parcels, i.e. postal services. However, in this dissertation an address is considered in the broader sense as the description of a location not only for postal delivery, but for *all* kinds of service delivery, ranging from "physical" services such as utility services (water, electricity, sewerage, etc.), billing, courier, goods delivery, and emergency dispatch; to more "abstract" services such as opening financial accounts, credit applications, tax collection, and land and property registration. Farvacque-Vitkovic *et al.* (2005) describe the importance of street addresses from the perspective of the general public, local governments and the private sector. This broader definition of an address is also found in Coetzee and Cooper (2007a) and Davis and Fonseca (2007). In this dissertation, any information about the recipient of the service delivery (whether a person or an organization) is excluded from the address.

Address data refers to a collection of addresses, and *reference data* to data according to which other information can be referenced unambiguously. Thus, *address reference data* is a collection of addresses according to which other information can be referenced unambiguously. This dissertation is about address reference data but for simplicity reasons the term 'address data' is used when referring to 'address reference data'.

As an example of the use of an address as a reference, in a corporate database a customer could be referenced to an address, or in a disaster management situation an incident could be referenced to an address. Refer to Figure 4. This implies that the address is an independent entity, object or feature to which the other information is linked. This idea of an address as an independent object to which other data entities are linked, is in stark contrast to the way in which an address is stored in many current corporate and other databases, namely as a number of free text attributes of the data entity, which are extremely difficult to verify or quality check. When it comes to address data sharing and exchange, it is difficult to compare these free text attributes in order to, for example, find duplicate addresses.

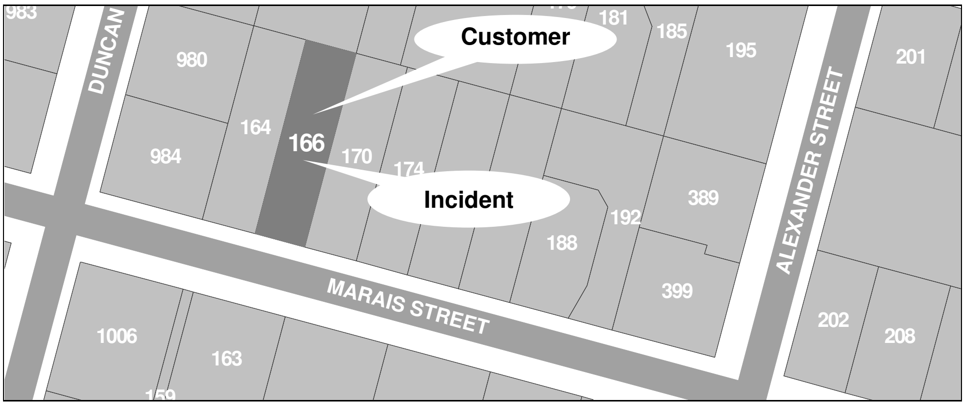

Figure 4. Address data

The idea of an address as independent object suggests that an address reference dataset comprises independent address features. An address reference dataset would thus be a collection of individual address features, where a feature according to Cooper (1993) is described as a uniquely identifiable set of one or more objects in the real or potential world where the defined characteristics of the objects are consistent throughout all the objects. The importance of address data as reference data is confirmed in the preparatory work of the European program for an SDI, INfrastructure for SPatial InfoRmation in Europe (INSPIRE), where the concept of 'reference data' has been defined as a category of datasets that plays a special role in the infrastructure. According to the INSPIRE definition (Rase *et al.* 2002), reference data must fulfill the following three functional requirements:

- provide an unambiguous location for a user's information;
- enable the merging of data from various sources; and
- provide a context to allow others to better understand the information that is being presented.

Addresses fulfill all three of these requirements. In numerous legacy and modern IT systems, address information is recorded with the purpose of having an unambiguous identification of the real estate, customer, citizen, business or utility entity in question. Secondly, addresses are used as one of the most important mechanisms to merge or link information from different sources together, e.g. when a bank uses the customer's address to look up information on real estate or insurance. Thirdly, addresses are used every day by citizens, businesses and government as a human understandable description of the location of a specific piece of information; for example, the address label on letters or goods for delivery is meant to give every actor in the delivery process a clear understanding of the desired final destination. As a result of these considerations, addresses have been included explicitly in 'Annex 1' of the final INSPIRE Directive that lists the priority spatial reference datasets (Directive 2007/2/EC of the European Parliament 2007).

Compartimos, the reference model for an address data grid, which is presented in Chapter 4 is designed for an address dataset of which the features can be used as references for all kinds of other information, such as those mentioned above, namely real estate, customers, citizens, businesses or utility entities.

An *addressing system* refers to the system according to which addresses are assigned. The US Draft Street Address Standard (2005) refers to an addressing system as an addressing scheme, also known as an address numbering system or an address numbering grid. A *spatial reference system* is a system for identifying position in the real world (ISO 19112:2003). Because an address identifies a position in the real world, the individual addresses in an addressing system can be regarded as locations in a spatial reference system (Coetzee *et al.* 2008b). According to ISO 19111:2006, *Geographic information – Spatial referencing by coordinates* and ISO 19112:2003, *Geographic information – Spatial referencing by geographic identifiers*, spatial references fall into two categories:

1. those using coordinates, and
2. those using geographic identifiers.

Coetzee *et al.* (2008b) identified a third type of spatial reference system, the linear reference system as defined in ISO 19116:2004, *Geographic information – Positioning services*, which identifies a location by reference to a segment of a linear geographic feature and distance along that segment from a given point. Theoretically, an addressing system can be regarded as any one of the three types of spatial reference systems:

1. A *coordinate reference system* is a coordinate system that is related to the Earth by a datum (ISO 19111:2007), i.e. location is specified by reference to a datum. For example, the WGS84 Latitude and Longitude coordinates of *(25°45'20.90", 28°13'56.98")* specify the location of the centre point of the IT building on the main campus of the University of Pretoria by reference to the World Geodetic System 1984 ellipsoid, commonly known as WGS84, with coordinates of the Hartebeesthoek Radio Astronomy Telescope used as the origin of this system (Chief Directorate: Surveys and Mapping 2004). While such a location in an urban area usually has one or more equivalent human-understandable address, in rural areas where village and street names are not yet formalized, a coordinate is sometimes the only reference to a dwelling and thus in a way, constitutes an address. In this case the addressing system is a coordinate reference system.

2. A *linear reference system* specifies the location by reference to a segment of a linear geographic feature and distance along that segment from a given point (ISO 19116:2004). '200m West of the filling station along Burnett Street' is an example of a linear reference where 'Burnett Street' is the linear geographic feature and '200m West' is the distance from the given point, the filling station. In some addressing systems, addresses are linear references and then, for example, *'310 King Street'* specifies the following location: proceed 310 meters (distance) along King Street (linear geographic feature) from its origin (given point). Thus, in this case the addressing system is a linear reference system. The Australian rural addressing system (AS/NZS: 4819:2003) is an example of an addressing system that is a linear reference system.

3. A *geographic identifier reference system* is a system for identifying position in the real world based on geographic identifiers, i.e. labels or codes, that identifies location (ISO 19112:2003). These reference systems tend to be based on hierarchies of geographic identifiers that identify, with increasing accuracy, a position in the real world. For example, *Country>Province>Municipality>Suburb* is an example of a South African geographic identifier reference system, and *South Africa>Gauteng>City of Tshwane Metropolitan Municipality>Hatfield* specifies a location according to this system. This geographic identifier reference system can be extended to include street names and street numbers, as in

Country>Province>Municipality>Suburb>Street>Street Number, thus a typical street address such as *South Africa>Gauteng>City of Tshwane Metropolitan Municipality>Hatfield>Pretorius Street>1083* specifies a location according to this system. In this case the addressing system is a geographic identifier reference system.

In the Compartimos address data model that is presented in Chapter 4 each address is linked to an addressing system, specifying the content and structure of the address.

2.2.2 Production, maintenance and distribution

Due to their service, infrastructure and land administration responsibilities, it is commonly found that a local authority establishes and maintains address data for its area of jurisdiction (Levoleger and Corbin 2005, Williamson *et al.* 2005, Coetzee *et al.* 2008b). The assignment and maintenance of addresses is usually closely linked to the responsibilities of the local authority and therefore focuses on fulfilling the local authority's requirements for service delivery. Unless there is guidance for address assignment and maintenance, either through legislation or through some coordinating body with a mandate, each local authority tends to apply its own rules in terms of addressing systems, naming conventions, record of address history, positioning of the coordinate in relation to the land parcel, etc. These differences at local authorities introduce syntactic and semantic heterogeneities that are a challenge when address data is collated for an area that spans the jurisdictions of multiple local authorities.

The production and maintenance of address data is, however, not necessarily limited to local authorities; for example, almost all addresses in rural areas of South Africa have been assigned nationally by the South African Post Office, Statistics South Africa, national departments, national utilities and private companies (Coetzee *et al.* 2008b). Table 3 lists some of the address data producers in South Africa. The wide range of purposes for which address data is produced results in many different formats and models of address data. Some of these organizations have allocated addresses according to their own individual addressing systems and have painted their corresponding address number on house doors, particularly in rural areas, resulting in a single house with three different numbers on its door (Coetzee and Cooper 2007b). Multiple independent organizations assigning address data for a variety of purposes increase the challenge of syntactic and semantic address data heterogeneity. Belussi *et al.* (2006) identify another heterogeneity factor due to multiple data providers, i.e. the different levels of accuracy at which spatial data is captured due to data providers employing different update processes. Coetzee and Cooper (2008) show how some of this heterogeneity can be overcome by implementing the South African address standard at local municipalities in South Africa.

Table 3. Address data producers in South Africa (Coetzee and Bishop, 2008)

Source	Type of data	Purpose	Typical Coverage	Formats
Town planning departments at municipalities	Land parcels and their assigned street names and numbers	Support function to other municipal departments	Municipality	Paper maps, CAD drawings, or GIS databases
Property valuation rolls at municipalities	Property description (as per deeds registry) together with a postal address	Property Valuation	Municipality	Paper printouts
Consulting town planners	Plan showing the layout of proposed erven and their assigned street names and numbers for new development	Town Planning	Town or suburb	Paper maps, CAD drawings, or GIS databases
South African Post Office (SAPO)	A list of SAPO-approved place names with their postcodes. No spatial information included	Postal mail delivery	National	Comma delimited text file
Statistics South Africa	Database of coordinates for dwelling locations, sometimes with an address	Household surveys	Per area as required for a survey	Proprietary GIS databases
Telephone and electricity utilities	Service delivery points and/or dwellings with GPS coordinates and custom addresses	Support planning and deployment of services	National	Proprietary GIS databases
State IT Agency (SITA)	Address data sourced from a single private company	Provide data and services to government departments only	National	Proprietary GIS databases
Private Companies (non-spatial)	Compiled from the customer databases of various organizations; often includes the name of an individual or business	Direct marketing	Provincial, National	Relational database tables or comma delimited text files
Private Initiatives (spatial)	Source address data from data producers listed above, and aggregate them into a national database	Address-related service provision, either by the company itself or sold to a third party	National	GIS database formats

Most service delivery related work done in a national government department or a commercial address-related service to a larger community requires address data for an area that spans multiple jurisdictions. If address data is produced at individual local authorities and/or other independent organizations, this implies that data has to be collated from the different sources of address data. The collation can be done either dynamically or at regular intervals. Dynamic collation has the advantage of being able to provide the latest up-to-date data but there is usually a penalty on the response time for fetching the data on the fly. On the other hand, collating the data at regular intervals holds the advantage that data can be cleaned and indexed when received at those intervals, resulting in shorter access response times but the disadvantage is that the data is only as current as the latest collation interval, which in practice ranges between three and six months, as can be seen from the G-NAF in Australia (Paull 2003), the GeoDirectory of Ireland (Fahey and Finch 2005) the AfriGIS data release cycle (AfriGIS 2008). In developing countries where address data is in flux, such an interval is problematic.

A European survey on addresses and address data (Levoleger and Corbin 2005) gives clear evidence that address systems with a long history, along with address master files or address registers, exist in many European countries. Some of these address registers or master files are collated from individual local authorities, such as in the Netherlands, Norway, Austria and the National Land and Property Gazetteer (NLPG) in the UK; while others are produced on a national scale such as the GeoDirectory in Ireland and the AddressPoint dataset produced by the Ordnance Survey in the UK. There are, however, also European countries where address data is maintained at local authorities and not (yet) collated into a national dataset, such as Croatia, Portugal, Germany, France, and Hungary.

In Australia the Public Sector Mapping Agencies (PSMA) follow a semi-automated process of massaging contributor address data from various agencies and organizations into the standard format of the Geocoded National Address File (G-NAF®), which is distributed quarterly (Paull 2003). In developing countries such as Brazil and India such comprehensive databases of address data are usually not readily available. The large cities in these countries often contain slums, shantytowns, and other types of low-income areas that are characterized by irregular occupation, and often in these areas there are neither street signs nor individual address signs at each dwelling. Also, in many cases the addressing database is not as complete as it should be, due to lack of information or to the cost of generating and maintaining a detailed database in places where fast and chaotic growth, and irregular land occupation, are predominant (Davis and Fonseca 2007). In South Africa, a developing country, there is currently not a public sector initiative for a national address dataset, but a number of private sector companies, including AfriGIS (www.afrigis.co.za) and Knowledge Factory (www.knowledgefactory.co.za), have produced national address datasets that are compiled from local authority datasets and released quarterly.

Compartimos, the reference model presented in this dissertation, accommodates the reality of dynamic, distributed, uncoordinated and diverse production and maintenance of address data, thereby allowing for the dynamic collation and distribution of heterogeneous address data sources from individual local authorities and other relevant organizations.

2.3 Spatial Data Infrastructure (SDI)

2.3.1 The concept

An SDI involves everything and anything that is required to make spatial data from various sources useful and available to as wide an audience as possible. The list of components of an SDI varies in literature but generally includes *spatial data*, technologies, systems (hardware and software), standards, policies, legislation, agreements, human and economic resources, institutions, organizational aspects and *people* (US Executive Order 1994, Jacoby 2002, Crompvoets *et al.* 2004, Georgiadou *et al.* 2005, De Man 2006, Rajabifard *et al.* 2006, Masser *et al.* 2007). In fact, an SDI aims to make *spatial data* usable by *people*, and the technologies, systems (hardware and software), standards, policies, legislation, agreements, human and economic resources, institutions, and organizational aspects have to be carefully orchestrated to make this possible. See Figure 5.

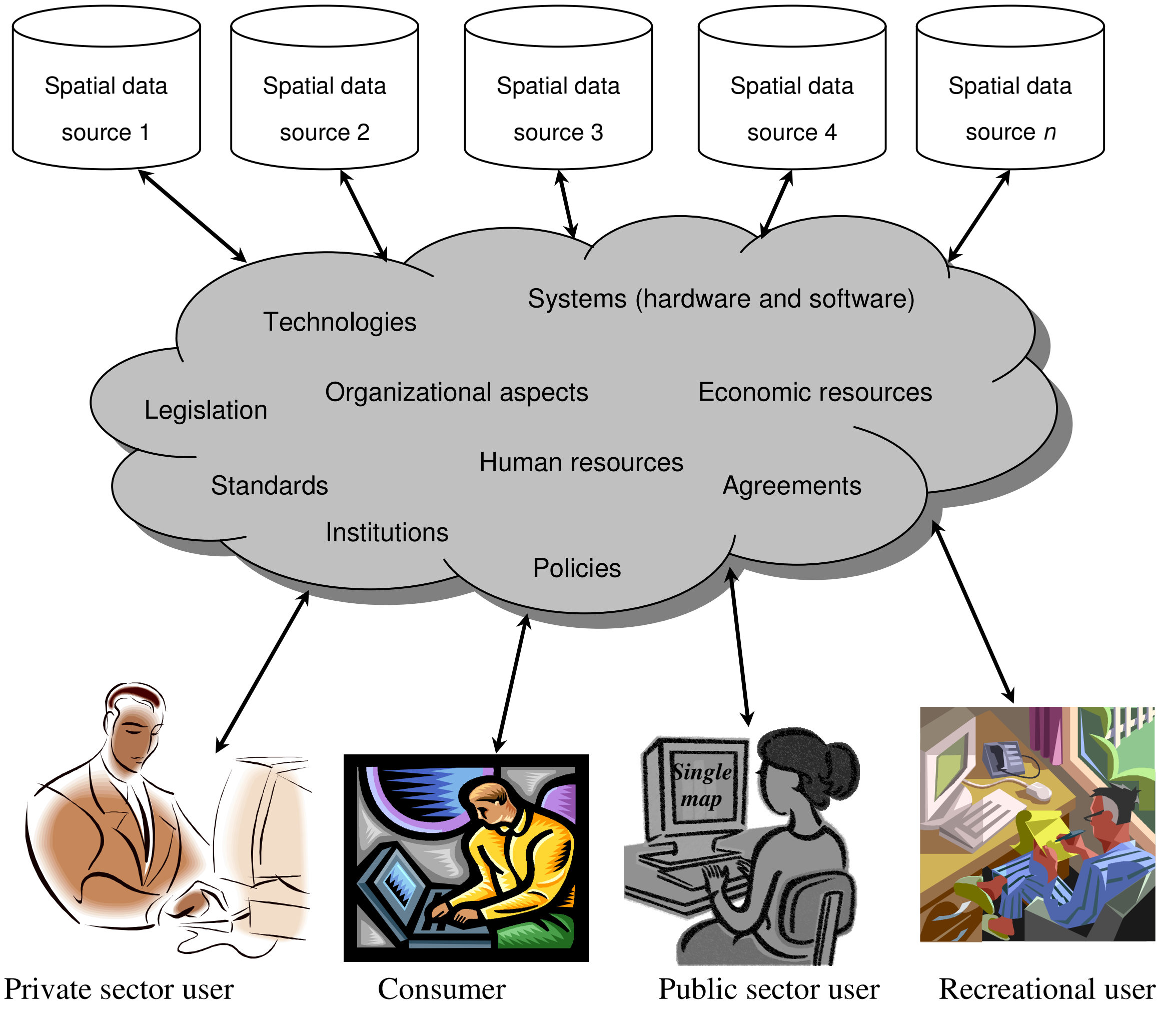

Figure 5. An SDI aims to make multi-source spatial data usable by people

2.3.2 Origins and current reality

SDIs emerged from the 1980s when countries such as the USA and Australia, for example, started to develop data access relationships, which became the precursor to the development of national SDI initiatives. At this time, countries developing SDIs had limited knowledge about different dimensions and issues of SDIs, and rather less experience of such development. Each country designed and developed an SDI based on their specific requirements and priorities and nationally specific characteristics. However, these early initiatives provided documentation of researchers' and practitioners' experiences along with status reports of SDI initiatives, thereby establishing a knowledgebase from which to learn and thus develop and adjust existing SDI initiatives as well as design and plan new SDI initiatives (Crompvoets *et al.* 2004).

Since those early days many countries have started planning and implementing spatial data infrastructures and the knowledgebase is ever increasing, with reports on these initiatives in both journals, for example, Jacoby on Australia (2002) and Georgiadou *et al.* on India (2005), as well as on conferences, such as the North American Urban and Regional Information Systems Association (URISA) conference, the Australasian Urban and Regional Information Systems Association (AURISA) and the conference held by the Global Spatial Data Infrastructure Association (GSDI), of which the following are examples: Iglesias on Chile's SDI (2008), Wytzisk *et al.* on the GDI-DE in Germany (2008), de Bree *et al.* on the Dutch SDI (2008); Valentin and Cabello on SDI initiatives in Spain (2008) and Jin-Hsiang and Chung-Chi on the geospatial one-stop in Taiwan (2008).

	1st generation SDI	**2nd generation SDI**	**3rd generation SDI**
Approach	Product-based	Process-based	Uncoordinated decentralized activity
Focus	Data production, database creation and centralizations	Use and application of data, Web services	A problem-oriented virtual world to facilitate decision making
Key driver	Data	Users and their needs	Decision making
Role players			
- National authorities	Strategic and operational		Strategic but less important
- Local authorities	Operational to lesser degree		Operational
- Private sector	Not involved		Operational

Figure 6. SDI evolution, adapted from Rajabifard *et al.* (2006)

Rajabifard *et al.* (2006) identify three generations of SDIs. Refer to Figure 6. The *first generation* that emerged from the 1980s mostly followed a top-down *product-based* approach. In these early SDIs, national mapping agencies played a major strategic and operational role. The product-based SDI model tends to be data-producer- and national-mapping-agency-led, focusing on data production, database creation, and centralization. In the first generation SDI, *data* used to be the key driver.

Around the year 2000 a transition to the *second generation* SDIs occurred when leading SDI initiatives started to take advantage of the capabilities of the Internet and the World Wide Web. The focus shifted to the creation of an infrastructure to facilitate the management of information access instead of the linkage to existing and future databases, and the development model changed from being product-based to a more *process-based* approach. Data sharing drives the process-based SDI model, as well as re-using data collected by a wide range of agencies for a great diversity of purposes. This model also sees the trend of moving away from the centralized structures of most early SDIs to the decentralized and distributed networks that are a basic feature of the Internet and World Wide Web (Rajabifard *et al.* 2003). In the second generation SDI, the *users and their needs* are the key drivers and consequently the focus shifted from the data in itself to the use and application of data, including the introduction of Web services for providing data access. Web services are regarded as the main technological indicator of a second generation SDI.

Initial SDI development was the domain of national governments whose role it was to map and collect small-scale data about a nation. They played both a strategic and an operational role in SDI development, following a top-down approach to policy development. The building of the infrastructure was seen as a national role, especially within developing countries whose sub-national or local level of government is generally not as well developed as that of developed countries. Naturally, the involvement of local governments and the private sector was not as coordinated as that of a national government with resulting uncoordinated SDI activity. When policy development came from the national level, there was no real driving force for the other two sectors to play in SDI development. The *third generation* SDI is currently in the making, where these roles are changing.

Current trends and development within SDIs have shown that the roles of the three major players – national governments, local governments and the private sector – are changing. The previous influence of national governments at both the strategic and the operational level has diminished, although there is still a strong case for a strategic national government role in SDI through coordination, as is evident from the European program for an SDI, INSPIRE (Directive 2007/2/EC of the European Parliament 2007). The operational level of SDI that, in the first generation, was undertaken by national governments has now moved to the local government level. The involvement of the private sector has also grown substantially to the point where they are beginning to utilize, create, maintain, and influence the implementation of SDIs. This sets the stage for an *uncoordinated environment that is not subject to centralized control.* Harvey and Tulloch (2006) report that the successful establishment of large scale SDI datasets from the collation of local government datasets is not common and that a decentralized *federation-by-accord* data sharing model, although difficult to establish, is more sustainable in the long run. Craglia *et al.* (2008) confirm that the nature of more recent SDIs has changed with an increased umber of stakeholder organizations engaged in the process.

The scale and complexity of this uncoordinated activity in countries with a large land mass, large population, and heavily decentralized governance structure, such as the United States, is massive, given that more than 80 000 public bodies alone are involved in some way. This task is made even more difficult by a governance model that is based largely on consensus building and the extent to which coordination bodies such as, for example, the Federal Geographic Data Committee (FGDC) in the United States, the Spatial Information Council (ANZLIC) of Australia and New Zealand and the South African Bureau of Standards (SABS) in South Africa, lack the powers to enforce their standards or to impose sanctions on unwilling participants.

Due to the added number of SDI participants resulting from increased operational involvement of local government and the private sector, the heterogeneity of all aspects of the data has equally grown. This, together with the increased demand for spatial data resulting from the use of geobrowers such as Google Earth, NASA Worldwind, Microsoft VirtualEarth, as well as state-of-the-art technology such as GPS devices, and in-vehicle navigation, effected a similar *increase in the demand for data sharing* which SDIs have to somehow meet. Craglia *et al.* (2008) report that there is stronger emphasis on distributed data and processes, and the interoperability of services to discover, view, access, and integrate spatial information.

In summary, this new generation SDI, that Rajabifard *et al.* (2006) call the third generation SDI and Craglia *et al.* (2008) the next-generation SDI, faces some challenges:

- huge increases in the number of independent SDI stakeholders resulting in uncoordinated activity, and less strategic and operational activity by national and local authorities;
- increased heterogeneity of all aspects of the data; and
- an exponential increase in the demand for spatial data.

Craglia *et al.* (2008) identified a number of research priorities for the realization of next-generation SDIs. One of these is the integration of information from multiple heterogeneous sources, comprising data that is multi-disciplinary, multi-temporal, multi-resolution, multimedia and multi-lingual, requiring a multi-disciplinary approach. Also a research priority is computational infrastructures that can achieve integration of multiple systems delivering data, information and models in real-time from multiple sources. In this dissertation the use of a data grid – a scalable distributed architecture that functions without centralized control – in an SDI is analyzed. This analysis thus investigates the usefulness of a data grid approach for next-generation SDIs.

2.3.3 The future

While the second generation SDI was developed with the aim to facilitate access and sharing of spatial data hosted in a distributed environment, in the currently emerging third generation SDI users require precise spatial information in real time about real-world objects, together with the ability to develop and implement cross-jurisdictional and interagency solutions to public priorities such as emergency management; natural-resource management; water rights; and animal, pest, and disease control. In order to achieve this, the concept of an SDI is moving to *a new business paradigm, where SDI is the enabling platform* to promote the partnership of spatial-information organizations (public/private) to provide access to a wider scope of data and services, of a size and complexity that are beyond an individual organization's capacity. SDI as an enabling platform can be viewed as an infrastructure linking people to data through linking data users and providers on the basis of the common goal of data sharing (Masser *et al.* 2007).

According to Rajabifard *et al.* (2005) the technical basis for delivery of the enabling platform should be through an interoperability architecture based on *distributed*, custodial data management and *open* standards. The aim of this architecture is to allow initiatives to grow in an open environment that gives agencies the ability to operate in an integrated manner. The ability to deliver the concept of a spatially enabled platform, however, will also require an investigation of the *way in which that data will be stored in the future*. One of the key objectives of an SDI is to facilitate the interoperable environment through the ability to integrate multi-source datasets. New database-management software and technology promise to change both the way in which data are stored, as well as the underlying technology for the enabling platform in general. The benefits of such technology are already being seen in the concept of virtual libraries, emerging Grid computing technologies and super servers, as well as cloud computing where data and processing resources are managed on remote servers accessed over the Internet (Craglia *et al.* 2008).

What should be researched today is technology that can provide access to a wider scope of data and services, of a size and complexity that are beyond an individual organization's capacity and that can provide the enabling platform to realize the common goal of data sharing. As mentioned by Rajabifard *et al.* (2005), emerging Grid computing technologies hold the promise of changing both the way in which data is stored as well as the underlying technology for distributed architectures.

Compartimos, the reference model that is presented in this dissertation, shows how to grid-enable access to a wider scope of data and services that are beyond an individual organization's capacity, thereby realizing the common goal of data sharing. Thus Compartimos contributes to the future of SDIs.

2.4 Address data in an SDI

The typical responsibilities of local governments cause them to often become the custodians of street address and other land related data in a country (Williamson *et al.* 2005). The challenge that faces many countries is the establishment of national datasets from these numerous local datasets, or in the case of Europe to establish an international dataset from the numerous national datasets. The fact that address data is usually maintained on a local level but required on a wider scale implies that the principles of SDIs apply for collating address data into databases for national and international SDI, and making them available to as wide an audience as possible.

Moreover, address data forms one of the basic building blocks of an SDI, as can be seen from the fact that address data is included as one of the nine priority spatial reference dataset in 'Annex 1' of the European INSPIRE Directive (Directive 2007/2/EC of the European Parliament 2007).

Some of the implementations of address databases on a national scale such as those in Australia, the UK and Ireland follow the data-harvesting model where all local data is loaded into a single centralized database and published periodically. These initiatives are described in Jacoby *et al.* (2002) and McDougall *et al.* (2005) for Australia, by Morad (2002) for the UK, and by Fahey and Finch (2006) for Ireland. However, Harvey and Tulloch (2006) point out that due to a number of reasons the successful establishment of national datasets from the collation of local government datasets is not common. Their research into local government data sharing provided an evaluation of the foundations of spatial data infrastructures and indicated that a decentralized "federation-by-accord" data sharing model seems to be more sustainable, and thus, it seems there is a need to explore information architectures that support this "federation-by-accord" data sharing model.

Address data is an important dataset in a national or international SDI. The emerging concept of an SDI as the enabling platform that provides access to a wider scope of data and services, of a size and complexity that are beyond the capacity of individual organizations (as described by Rajabifard *et al.* 2006) is closely related to the concept of a grid as the enabling platform for providing flexible, secure, coordinated resource sharing among collections of individuals, institutions and resources spanning multiple administrative domains (as defined by Foster and Kesselman 1999). Thus there are some similarities between future SDIs and data grids. Coetzee and Bishop (2008) explore these similarities further in the context of address databases for national SDI in a paper that is included as Chapter 6 of this dissertation. This dissertation presents Compartimos, a reference model for the novel data grid approach to making address data in an SDI available on a national scale. This is a novel alternative to the centralized database approach and is in line with the sustainable "federation-by-accord" data-sharing model proposed by Harvey and Tulloch (2006), as well as the requirements for future SDIs described by Rajabifard *et al.* (2006) and Craglia *et al.* (2008).

2.5 Standards and technologies for address data in an SDI

An SDI aims to make *spatial data* usable by *people*, and the technologies, systems (hardware and software), standards, policies, agreements, human and economic resources, institutions, and organizational aspects have to be carefully orchestrated to make this possible. As discussed in Chapter 1 this dissertation gravitates towards the technical aspects of an SDI, i.e. the technologies, systems, standards and policies: refer to Figure 5. Agreements, human and economic resources, institutions, and organizational aspects are discussed on the periphery and only in relation to the technical aspects, i.e. either their impact on the technical aspects or how the technical aspects impact on them. A short overview of technologies and standards that are currently used to build actual SDI systems is therefore warranted. In the following paragraphs standards developed by the ISO/TC 211, *Geographic information/Geomatics* and the Open Geospatial Consortium (OGC) are discussed. These standards have become the cornerstone of most SDIs around the world (Craglia *et al.* 2008), and are therefore relevant to the work described in this dissertation.

To enable the design of an interoperable and interacting system in a heterogeneous environment, a guiding set of concepts and principles, sometimes referred to as a framework, along with actual standards, is required. In particular these standards have to provide for both content (the data itself), as well as functionality (accessing and updating the data). The ISO/TC 211 scope statement describes this standardization work (www.isotc211.org):

> *Standardization in the field of digital geographic information.*
>
> *This work aims to establish a structured set of standards for information concerning objects or phenomena that are directly or indirectly associated with a location relative to the Earth.*
>
> *These standards may specify, for geographic information, methods, tools and services for data management (including definition and description), acquiring, processing, analyzing, accessing, presenting and transferring such data in digital/electronic form between different users, systems and locations.*
>
> *The work shall link to appropriate standards for information technology and data where possible, and provide a framework for the development of sector-specific applications using geographic data.*

One of the first standards developed by ISO/TC 211, is the ISO 19101:2002*, Geographic information – Reference model,* which provides a framework for the 19100 series of standards, i.e. all other standards developed by ISO/TC 211. ISO/TC 211 has published a large number of standards, as well as reports, that are used in many SDI implementations around the world. Examples of standards that are especially relevant to SDIs include ISO 19111:2007*, Geographic information – Spatial referencing by coordinates* which describes the minimum data required to define 1-, 2- and 3-dimensional spatial coordinate reference systems; ISO 19115:2003*, Geographic information – Metadata,* which defines the schema required for describing geographic information and services, providing information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data; and ISO 19117:2005, *Geographic information – Portrayal* which provides a schema definition for the portrayal of geographic information in a form understandable by humans, including the methodology for describing symbols and mapping of the schema to an application schema.

The Open Geospatial Consortium, Inc.® (OGC) is a non-profit, international, voluntary consensus standards organization that develops standards for geospatial and location based services (www.opengeospatial.org). The OGC Reference Model (ORM) provides an architecture framework for the ongoing work of the OGC and describes this framework from the viewpoint of information (features), computation (services), engineering (architectures) and technology (platforms). OGC have come up with a number of specifications for Web services, including but not limited to the Web Catalogue Service (WCS), Web Feature Service (WFS), Web Processing Service (WPS) and the Web Map Service (WMS). These services can be combined to construct an SDI with a service-oriented architecture, and follow the trend of Web services that is commonly seen in second generation SDIs. The loosely coupled nature of service-oriented architectures makes them ideal for distributed and heterogeneous environments, as found in SDIs.

It should be noted that since OGC Web services were evolved in parallel with the evolution of the Web service standards by the World Wide Web Consortium (W3C) and the Organization for the Advancement of Structured Information Systems (OASIS), OGC Web services initially did not comply with the Web services standards from the W3C and OASIS, such as the standards for the Web Services Description Language (WSDL), the SOAP protocol and Universal Data Description Discovery and Integration (UDDI) (Zhao *et al.* 2007). One of the achievements of the recently completed OGC Web Services, Phase 5 (OWS-5) Testbed, an initiative of OGC's Interoperability Program, was the development of SOAP and WSDL interfaces for four services: WMS, WFS-T, WCS-T, and WPS (OGC 2008b), showing that OGC is paying attention to the requirement identified above.

ISO is an international organization and its members are mainly from the public sector, including national standards bodies and organizations. On the other hand, according to the OGC website at www.opengeospatial.org, only three of OGC's seven strategic members are from the public sector and most of it's fourteen principal members are from the private sector, including companies such as Oracle, Google, Microsoft, Intergraph, Bentley Systems, ESRI and Autodesk. In general, ISO has broader goals and is working at a level of abstraction above OGC so that the two efforts complement each other, and both are necessary. ISO's work is not likely to result in immediate implementation-level specifications, so it is in both organizations' mutual interest to see that OGC's implementation specifications fit into the ISO framework (Peng and Tsou 2003).

Cooperation between ISO/TC 211 and OGC through the Joint Advisory Group (JAG) ensures that standards are developed and published in a coordinated fashion. Both ISO/TC 211 and OGC also collaborate with other organizations, for example, ISO/TC 211 and the European Committee for Standardization (CEN), *CEN/TC 287, Geographic information*, have jointly developed a number of standards, and OGC and the OASIS recently announced progress on their standards collaboration after a Memorandum of Agreement was signed in 2006 (OGC 2008a). ISO - the whole organization, not a specific technical committee - and the Universal Postal Union (UPU) have agreed to increase their collaboration and will set up a contact committee of six officials responsible for implementing the provisions of the agreement (ISO 2008). The first meeting of the UPU-ISO Contact Committee has been scheduled for 18 November 2008 and addressing is on the agenda (Mathur 2008). This is of particular interest to address data standards, which are discussed in this dissertation.

Standards for address data have been developed and are currently being developed by a number of countries and international organizations. These include Australia and New Zealand (as a joint effort), Denmark, South Africa, the United Kingdom, the United States of America, the Universal Postal Union (UPU), the International Organization for Standardization (ISO) and the Organization for the Advancement of Structured Information Standards (OASIS). While the UPU standard (UPU S42 2006) narrowly focuses on postal addresses, and the OASIS standard on addresses for a party (customer or business) that can include geospatial coordinates, the national standards have tended to cater for all forms of service delivery over and above mere postal delivery and these national standards regard an address as a stand-alone independent geographic feature, in other words, the address is a reference (Coetzee *et al.* 2008b).

A European survey on addresses and address data (Levoleger and Corbin 2005) shows that although address systems exist in European countries, only very few published standards for address data exist, complicating the INSPIRE task of 'interoperable and seamlessly accessible' address data sets 'across all of Europe', and it is expected that a European address standard will have to be developed. Coetzee *et al.* (2008b) analyzed a number of standards and came up with some guidelines for a potential future international address standard:

- The standard should be an abstract standard, providing a framework for describing address systems across the world. A national or regional address standard could be produced as a profile (i.e. subset) to describe a very specific addressing system. An address (e.g. '1083 Pretorius Street, Hatfield, 0083') would be an instance of a particular profile.
- The standard should provide common terms and definitions of an address, address elements and related concepts.
- The standard should aim to make the address data from the multitude of addressing systems exchangeable.
- The standard should also provide a data model that enables the integration of address data based on multiple addressing systems.

A first attempt at an 'international' definition for an address is found in Cooper (2008). This definition is based on an analysis of definitions for an address found in existing address standards. Address data exchange in a single country like South Africa is described in Coetzee (2008), but in this dissertation address data exchange across international borders is proposed by means of an interoperable address data model to store and represent address data from different countries is presented as part of the Compartimos reference model. To illustrate the use of the data model, addressing systems described in the draft SANS 1883, *Geographic information - South African address standard* are presented in a data model based on ISO 19112:2003, *Geographic information - Spatial referencing by geographic identifiers* and ISO 19115:2003, *Geographic information – Metadata*. Lessons learnt from the Compartimos address data model could be valuable input into an international address data standard.

International geospatial standards and specifications for both data content and functionality to access and update the data are currently used in SDI implementations around the world. For Compartimos it is important that these standards from the geospatial community are considered and used where possible. On the data content side, various address standards exist and are successfully used in national SDIs of countries such as Australia and the UK. A current initiative in the ISO/TC 211 community with involvement from the UPU, INSPIRE and a number of countries considers issues related to an international address standard, and explores the feasibility of the development of an international geospatial address standard (Coetzee *et al.* 2008a, Cooper and Coetzee 2008). On the functionality side, standards and specifications follow the service-oriented approach by describing Web services. These standards are tightly coupled with the technologies that are used in the systems of an SDI so that, for example, OGC Web service specifications require, if not a fully service-oriented architecture, at least an approach that allows for service orientation. The relevance and applicability of existing geospatial standards, including address standards, to Compartimos is discussed in Chapter 5 in the section on technology choices.

2.6 Related Work

In this section research and implementations relating to address data in an SDI are described in order to illustrate the novel GISc aspects of the work in this dissertation. More related work in the Computer Science discipline is discussed in Chapter 3.

Chapter 6 provides an overview of address databases for national SDIs. In most of the examples reported in literature, address data is consolidated into a single centralized database that is distributed at regular intervals, and/or provided in online maps and/or made available through Web services (Paull 2003, www.nlpg.org.uk, www.adresse-info.dk). Where address-related Web services are provided, these can be integrated into other SDI systems and activities. In Denmark a business case report (National Survey and Cadastre 2005) analyzed the potential benefits of making the standard address identifiers (postcodes, street names, address numbers and coordinates etc.) accessible free of charge by means of a set of Web services which any IT developer could implement in Web applications or portals. The analysis concluded that the proposed Web services would improve the e-Government infrastructure by making standardized address data easy available for all sectors at a low cost and by reducing uncertainty and errors caused by wrong or imprecise address data. Within the first three years, it was estimated that the benefits would outnumber the costs by a factor of 12:1.

As reported in section 2.5 of this chapter, there is currently an initiative in the ISO/TC 211 community with involvement from the UPU, INSPIRE and a number of countries that considers issues related to an international address standard, and explores the feasibility of the development of an international geospatial address standard (Coetzee *et al.* 2008a). Another development in the international address standardization arena is the UPU-ISO Contact Committee, which will have its first meeting with addressing on the agenda in November 2008. The Compartimos interoperable address data model is highly relevant to these two initiatives. Also of interest is a proposal to adopt the object-oriented formalism of the ISO 19100 series of standards as a canonical data model (CDM), a data model that can be understood by all participating systems, for modeling interoperable geographic information bases and their applications (Jang and Kim 2006).

Compartimos deviates from the centralized database approach by creating a novel distributed data grid architecture in which address data is made available at its source, i.e. there is no physical consolidation into a centralized database. In contrast, Craglia *et al.* (2008) report that a future Digital Earth might be built using a computer system architecture in which data and processing resources are managed on remote servers accessed over the Internet, and which is now being referred to as 'cloud computing'. This would also be a distributed approach, albeit slightly different to the data grid approach.

First reports on Grid computing technologies in SDI environments are found, amongst others, in the papers by Zhao *et al.* (2004), Aloisio *et al.* (2005a), Shu *et al.* (2006), Wei *et al.* (2006) and Di *et al.* (2008), and the author expects that the recently initiated collaboration between OGC and the Open Grid Forum (OGF) (GridToday 2007) will start adding to the momentum. The recently launched GDI-Grid project and the Canadian Geospatial Data Infrastructure Interoperability Pilot are discussed in the Related Work section of Chapter 3. Other reports focus on geospatial processing (in contrast to geospatial data sharing) on a grid (Schaeffer and Baranski 2008, Lanig and Zipf 2008). Examples of reports on research on service-oriented architectures in relation to SDI are found in Granell *et al.* (2007), Liang *et al.* (2007), Brauner and Schaeffer (2008) and Molina and Bayarri (2008). Béjar *et al.* (2008) propose an architectural style, a pattern, for SDIs, which is defined under the component-and-connector architectural view type, extending the client-server and shared-data styles. The style was created after analyzing six of the most relevant SDIs and geo-service architectural proposals with the objective to capture, unify and systematize the previous knowledge on SDI architectural models. A comparison between this style and the data grid approach proposed in this dissertation would provide for an interesting analysis of the Compartimos reference model, which could be considered in future work. To date the author has not found any reports of address data sharing (in contrast to processing and spatial data in general) on data grids.

In summary thus, in relation to the GISc discipline the work in this dissertation is a novel approach to address data sharing in an SDI, and the work on the Compartimos interoperable address data model is extremely relevant at this point in time in light of the current initiative towards an international address standard.