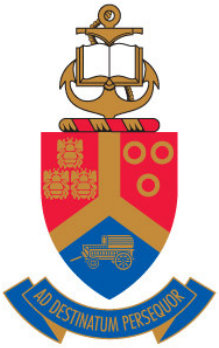


UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Department Computer Science

An analysis of a data grid approach for spatial data infrastructures

by

Serena Martha Coetzee

A dissertation submitted in partial fulfillment of the requirements for the degree of

Philosophiae Doctor (Computer Science)

in the

Faculty of Engineering, Built Environment and Information Technology,

University of Pretoria, Pretoria

November 2008

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Faculty of Engineering, Built Environment and Information Technology

Abstract

An analysis of a data grid approach for spatial data infrastructures

Candidate: Serena Martha Coetzee
Degree: PHD (Computer Science)
Supervisor: Professor Judith Bishop
Department: Computer Science

The concept of grid computing has permeated all areas of distributed computing, changing the way in which distributed systems are designed, developed and implemented. At the same time '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 collated from a vast number of different sources. A spatial data infrastructure (SDI) aims to make spatial data from multiple sources available to as wide an audience as possible. Current research indicates that, due to a number of reasons, data sharing in these SDIs is still not common.

This dissertation presents an analysis of the data grid approach for SDIs. Starting off, two imaginary scenarios spell out for the first time how data grids can be applied to enable the sharing of address data in an SDI. The work in this dissertation spans two disciplines: Computer Science (CS) and Geographic Information Science (GISc). A study of related work reveals that the data grid approach in SDIs is both a novel application for data grids (CS), as well as a novel technology in SDI environments (GISc), and this dissertation advances mutual understanding between the two disciplines. The novel evaluation framework for national address databases in an SDI is used to evaluate existing information federation models against the data grid approach. This evaluation, as well as an analysis of address data in an SDI, confirms that there are similarities between the data grid approach and the requirement for consolidated address data in an SDI. The evaluation further shows that where a large number of organizations are involved, such as for a national address database, and where there is a lack of a single organization tasked with the management of a national address database, the data grid is an attractive alternative to other models. The Compartimos (Spanish for 'we share') reference model was developed to identify the components with their capabilities and relationships that are required to grid-enable address data sharing in an SDI.

The definition of an address in the broader sense (i.e. not only for postal delivery), the notion of an address as a reference and the definition of an addressing system and its comparison to a spatial reference system contribute towards the understanding of what an address is. A novel address data model shows that it is possible to design a data model for sharing and exchange of address data, despite diverse addressing systems and without impacting on, or interfering with, local laws for address allocation. The analysis in this dissertation 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. In conclusion, results are presented and recommendations for future work, drawn from the experience on the work in this dissertation, are made.

Keywords: spatial data infrastructure, SDI, data grid, address data, addresses, data sharing, data exchange, grid computing, spatial data, geographic information, GIS, address standards, standards

Preamble

In the 1990s the major banks of South Africa joined forces to develop a national address dataset (NAD) for South Africa, which was later taken over by Media24, a member of the Naspers group. Around 2000, Naspers decided to narrow the focus of the Media24 group and AfriGIS became the new owners of the NAD. At that stage, I was a director at AfriGIS and gradually became involved in using, maintaining, expanding and marketing the 'AfriGIS NAD'. The NAD that we inherited from Media24 consisted of a number of files, each with a different set of attributes and no standard spatial orientation in relation to the cadastre: some NAD points were centered on the land parcel, while others were located at the street front. Our first task was to load all the NAD files into a single relational database with a uniform set of attributes. With the advent of our first big corporate clients, we published quarterly Release Notes with metadata about the dataset, as well documentation on the standard set of attributes of the AfriGIS NAD – a first experience in standardizing address data.

One benefit of having the NAD was that it enabled us to geocode address data on a national scale. The AfriGIS Intiendo address-matching toolset was developed to automate the geocoding process, using the AfriGIS NAD. Much of the Intiendo development took place in Dhaka, Bangladesh, by Reffat Zaman and his team of software developers, resulting in a globally distributed software development effort. Intiendo has been used successfully to geocode well-structured address data of quite a few corporate customers, each dataset comprising a few million address records each. However, the challenge of converting free format address data into structured address data that can be geocoded has never been solved to our satisfaction, mainly due to the fact that there are so many unknowns that cannot be solved without human intervention. For example, an address such as 'Arcadia 83 Pretoria' could refer to 'Arcadiastraat 83', the Afrikaans version of '83 Arcadia Street' or to 'Arcadia 0083 Pretoria', where 0083 is the postcode for the suburb of Arcadia. Another source of uncertainty are ambiguous suburb names, ambiguous suburb boundaries (the Intiendo solution to this ambiguity is described in Rahed, Coetzee and Rademeyer 2008), and incomplete addresses. These uncertainties are best resolved when addresses are captured into the system, and we were fortunate to assist some clients in developing user interfaces for address capturing – another experience in standardizing address data.

In order to offset the cost of maintaining the NAD we developed various value-added services, such as geocoding, address verification and routing, which we made available according to innovative business models with varying degrees of success. One of the biggest challenges in maintaining the NAD was to convert the address data from various formats into a standardized format that could be integrated into the national AfriGIS NAD. Over the years we saw a number of

failed attempts at developing a government initiated official NAD for the country. It seemed impossible to get so many stakeholders to agree and work together.

Against this background AfriGIS joined forces with Prof Judith Bishop from the University of Pretoria and submitted a proposal for a THRIP research project on 'Distributed Address Management' with the following project objectives:

Traditionally, national address databases have been built and maintained at a single central location, with a large computer and a single database. Centralized systems of this sort have inherent drawbacks: single point of failure, congestion, and low scalability. Given that address management in South Africa is in its infancy, the better approach would be to go for an incremental and multi-tiered system. The purpose of this project is to research, design and implement a prototype distributed spatial address database. The research questions will involve:

1. *whether grid is the correct way to go (we believe it is, but need to prove this to a largely untried and skeptical community of stakeholders);*
2. *how to design a multi-tiered approach to joining and accessing the grid taking into account changing levels of expertise and funding around the country, as well as the available bandwidth and connectivity.*

The THRIP project was approved and work on it started in January 2006. The work described in this dissertation was part of the project, which is jointly funded by the Department of Trade and Industry and AfriGIS. The paper published in the International Journal of GIS (IJGIS), which constitutes Chapter 6 of this dissertation, addresses the first objective. The second objective is addressed in the Compartimos reference model, presented in Chapters 4 and 5 of this dissertation.

At the same time as the THRIP project commenced, I took the initiative to lead the SABS project for the development of a South African address standard (SANS 1883), a project falling under the SC71E, *Geographic information* committee. The project was proposed and initiated in June 2004 but had not progressed since then due to a lack of resources. I arranged the first SANS 1883 project meeting in June 2006, and now at the closing of 2008, it is in the process of getting published by the SABS as a draft national standard for South Africa, the first locally developed standard by SC71E. My experience with the AfriGIS NAD proved invaluable to me on this project.

As a member of SC71E, I also became involved in its international mirror committee at the International Organization for Standardization (ISO), ISO/TC 211, *Geographic information/Geomatics,* where I recently took up the challenge of chairing the ISO/TC 211 Programme Maintenance Group (PMG). Antony Cooper from the CSIR is the convenor of working group 7 (WG) of ISO TC/211 and has a long history of involvement in GIS standards. He also

participated in the development of SANS 1883. We joined forces in publishing on address standards at conferences and in a journal.

My involvement in the South African address standard, SC71E and ISO/TC 211, as well as presentation at conferences as part of the THRIP project, brought me into contact with international role players on address standards. Together we have published papers and held workshops about address standards, and are exploring the possibility of an international address standard.

Two journal papers were published as a result of the above-mentioned THRIP project on Distributed Address Management. Coetzee and Bishop (2008) is included in this dissertation as Chapter 6; Coetzee and Cooper (2007) is included in Appendix E.

1. **Coetzee S** and Bishop J *(*2008). Address databases for national SDI: Comparing the novel data grid approach to data harvesting and federated databases, *International Journal of Geographic Information Science*, 26 September 2008, available online ahead of print edition at http://www.tandf.co.uk/journals/tf/13658816.html, accessed 26 October 2008.
2. **Coetzee S** and Cooper AK *(*2007b). What is an address in South Africa? *South African Journal of Science*, Nov/Dec 2007, **103**(11/12), pp449-458.

One project report under my supervision, as well as the following papers were presented at various conferences as part of the THRIP project (sorted alphabetically by author):

1. Acton D (2007). *Methods of charging for data in the NAD*, Hons project report, University of Pretoria, Pretoria, South Africa.
2. Arefin MA, Sadik MS**, Coetzee SM**, Bishop JM *(*2006). Alchemi vs Globus: a performance comparison, *4th International Conference on Electrical and Computer Engineering*, December 19-21 2006, Dhaka, Bangladesh.
3. **Coetzee S** *(*2008). Address data exchange in South Africa, *Proceedings of the ISO Workshop on address standards: Considering the issues related to an international address standard*, 25 May 2008, Copenhagen, Denmark.
4. **Coetzee S** and Cooper AK *(*2007a). The value of addresses to the economy, society and governance – a South African perspective, *45th Annual URISA Conference*, 20-23 August 2007, Washington DC, USA.
5. **Coetzee S** and Cooper AK (2008). Can the South African address standard (SANS 1883) work for small local municipalities? *Proceedings of the academic track of the 2008 FOSS4G Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape

Town, South Africa.

6. **Coetzee S**, Cooper AK, Lind M, McCart Wells M, Yurman SW, Wells E, Griffiths N and Nicholson MJL (2008). Towards an international address standard, *Proceedings of the GSDI-10 Conference*, Trinidad and Tobago, 25 – 29 February 2008.

7. Cooper AK and **Coetzee S** (2008). The South African address standard and initiatives towards an international address standard, *Proceedings of the academic track of the 2008 FOSS4G Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape Town, South Africa.

8. Rahed AA, **Coetzee S** and Rademeyer M (2008). A data model for efficient address data representation - Lessons learnt from the Intiendo address matching tool, *Proceedings of the academic track of the 2008 FOSS4G Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape Town, South Africa.

Acknowledgements

A posse ad esse ***From possibility to reality***

When I started work on my doctorate three years ago, I approached it like a project that has to be finished on time according to specification and within budget. I very soon discovered that a dissertation is more like creating a work of art that is admired for its beauty and not for its efficient and cost-effective production. If I could add up the costs of man-hours spent on this dissertation, any client would have fired me for over-expenditure long ago! Since I cannot repay in monetary terms all of you who have contributed, you will have to be content with an acknowledgement in my dissertation.

First of all, I would like to thank my supervisor, *Professor Judith Bishop*, who planted the seeds for this dissertation more than ten years ago, one night, late, after a dinner in Germany. It is a pleasure to work with someone who sets such high standards for herself and those working with her. Judith's advice on everything from writing papers to selecting conferences to life as an academic, have helped me find my way in this world of academia that seemed so very strange to me three years ago. She has also been a role model for achieving balance between career and being a mother.

Long, long ago in the days before cell phones and the Internet when the new South Africa was still in its infancy, I was privileged to join the development team of the first Windows GIS, *ReGIS*. Fresh from university, I learnt the art and skill of programming from 'real' programmers (men) like *Alf Tilley* and *Anton Koen,* and everything GIS-related from the remaining ReGIS team who are today scattered all over the globe: *Leon Jansen, Andreas Liebenberg, Felix du Plessis, Eugene Maré (in memoriam), Rudi Breedt, Wilhelm Herbst, Marianne Grobbelaar* and *Colin Hobson;* and of course the boss, *Johan Poolman*. These early experiences were a solid foundation for my understanding of spatial data, and established my passion for GIS.

Apart from *AfriGIS* generous funding of the THRIP Distributed Address Management project, I would like to thank *Magnus Rademeyer* and *Charl Fouché,* my co-directors from past AfriGIS days for their support and understanding for this venture of mine. The work in this dissertation originates from Magnus' visionary instinct to buy the NAD from Naspers and to start developing geocoding software. In my years at AfriGIS, the pioneer work on the AfriGIS datasets was a steady learning curve on the way to this dissertation, to which everyone who worked with me on the data team contributed: *Martha Burger, Marna Roos, Johan (JP) Roos, Johan (Vere) Nortjé, Hernand de Beer, Zanele Mkhomazi, Alwyn Esterhuizen, Dineke Vink, Pieter Geldenhuys, Magda Sandilands*, *Leentjie Reyneke, Adri Benadé, Sanli McSeveney, Martin van der Linde, Vino Naidoo, Mari Knoetze* and *Alwyn Moolman,* for that very first AfriGIS NAD data model. Similarly, the Intiendo team have and still are working studiously on understanding and improving geocoding efficiency: *Reffat Zaman, Ali*

Akter, Abdullah Al Rahed, Christopher Ueckermann, *Shibley Sadik* and *Iaan Roux*. And lastly, thanks to all the clients who entrusted us with their customer databases to geocode: this was the best way to get to know the South African address!

The work on this dissertation was supported in part by the THRIP project on Distributed Address Management (Application Reference TP2007081800001), jointly funded by the *South African Department of Trade and Industry* (dti) and *AfriGIS* (Pty) Ltd. I am grateful to the *Department of Computer Science* at the *University of Pretoria* for the opportunity to become a colleague, which has helped me understand the world of research and academia, without which I would not have been able to complete this dissertation.

The South *African Bureau of Standards (SABS)* initiated the project for the South African address standard (SANS 1883) on which I reported with my co-authors in various publications. Delegates from numerous organizations (too many to mention names!) actively contributed and participated in the SANS 1883 project meetings and workshops where the standard was developed. The funds from the award by the Small Grants Program of the *Global Spatial Data Infrastructure Association (GSDI)* helped to create awareness of SANS 1883 and facilitated attendance of delegates at SANS 1883. Leading the SANS 1883 project was an extremely rewarding exercise where I could plough some of my knowledge and experience back into the community, and I sincerely hope that the South African address standard will have a positive impact on South African citizens. Through the SABS I had the opportunity to attend plenary meetings of the ISO/TC 211*, Geographic information/Geomatics* where *Antony Cooper* and *Garth Mackway-Wilson,* through many discussions introduced me to the world of international standards. Finally, the exchange of ideas and experiences with international 'address' colleagues has helped shape the work in this dissertation: *Morten Lind, Randy Fusaro, Chris Corbin, Martha McCart-Wells, Ed Wells, Michael Nicholson, Nick Griffiths, Rob Walker, Bob Barr, Piotr Piotrowski, Joe Lubenow, Andrew Coote, Ram Kumar, Carl Anderson, Sara Yurman* and *John Hockaday*.

A special word of thanks to *Antony Cooper* for the time he spent reading a draft of this dissertation, producing invaluable comments in a very short timeframe. I hope to be able to return the favor one day (soon).

Pater et Mater, vobis gratias maximas ago pro eduactione meo atque pro exemplo vestro optimo. Vitae vestrae demonstrant quomodo homini pro communitate est vivendum. Et pater olim dixit: vetustas non est impedimentum discendi!

To my children, Cara and Max, who, indirectly, are responsible for my diverging on a journey into academia resulting in this work of art: the stress-free time that I now spend with you is the best reward! And most of all, to my husband, Jan, for our mutual understanding of '*My life makes your life possible*'.

Alea iacta est… ***The die is cast…***

Table of Contents

List of Tables

List of Figures

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.

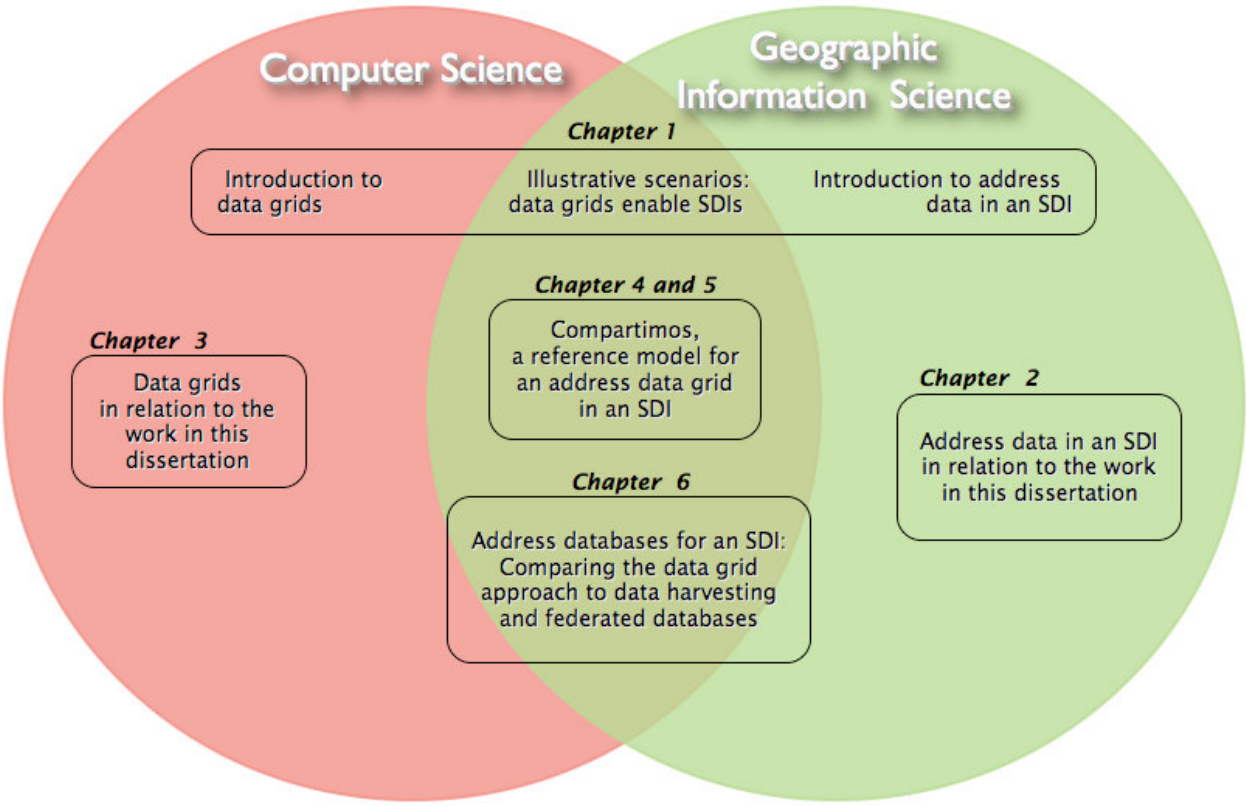


Figure 1. The chapters in this dissertation in relation to CS and GISc

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.

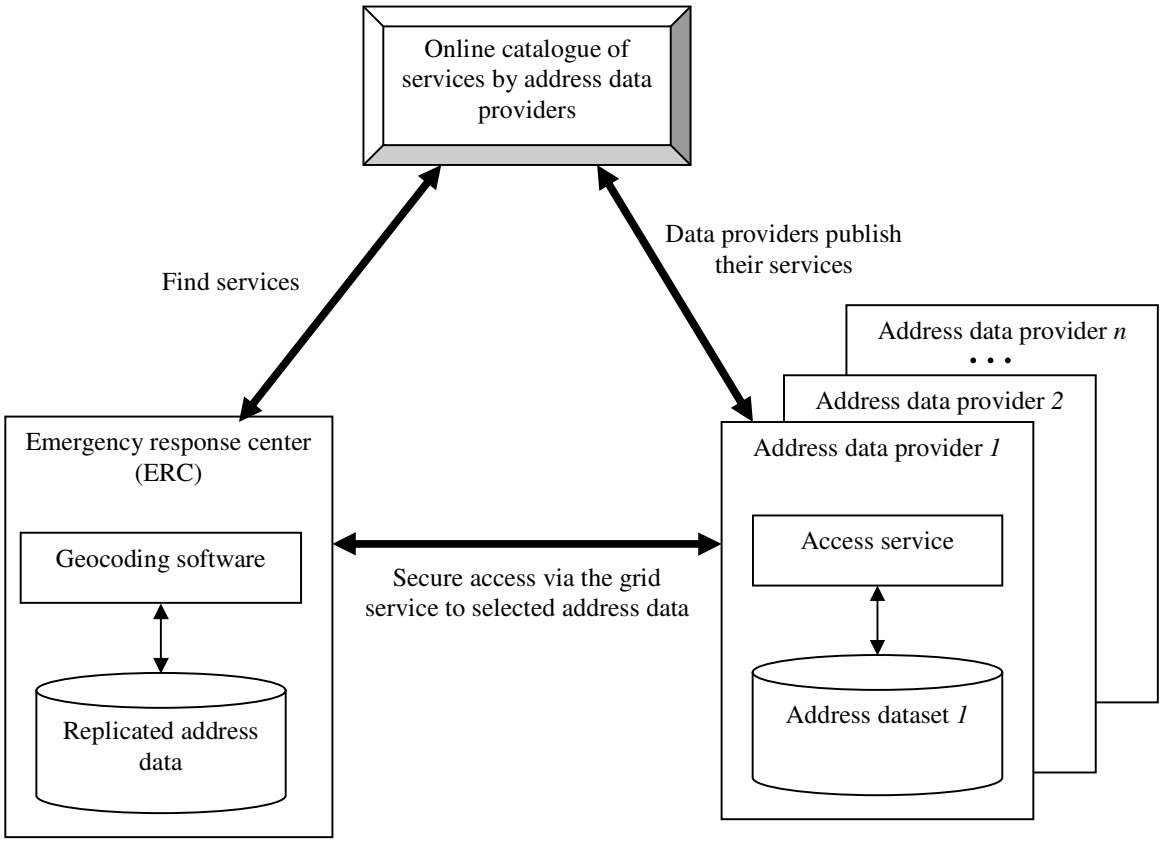


Figure 2. Scenario 1 – Mapping the locations of damage and distress reports

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.

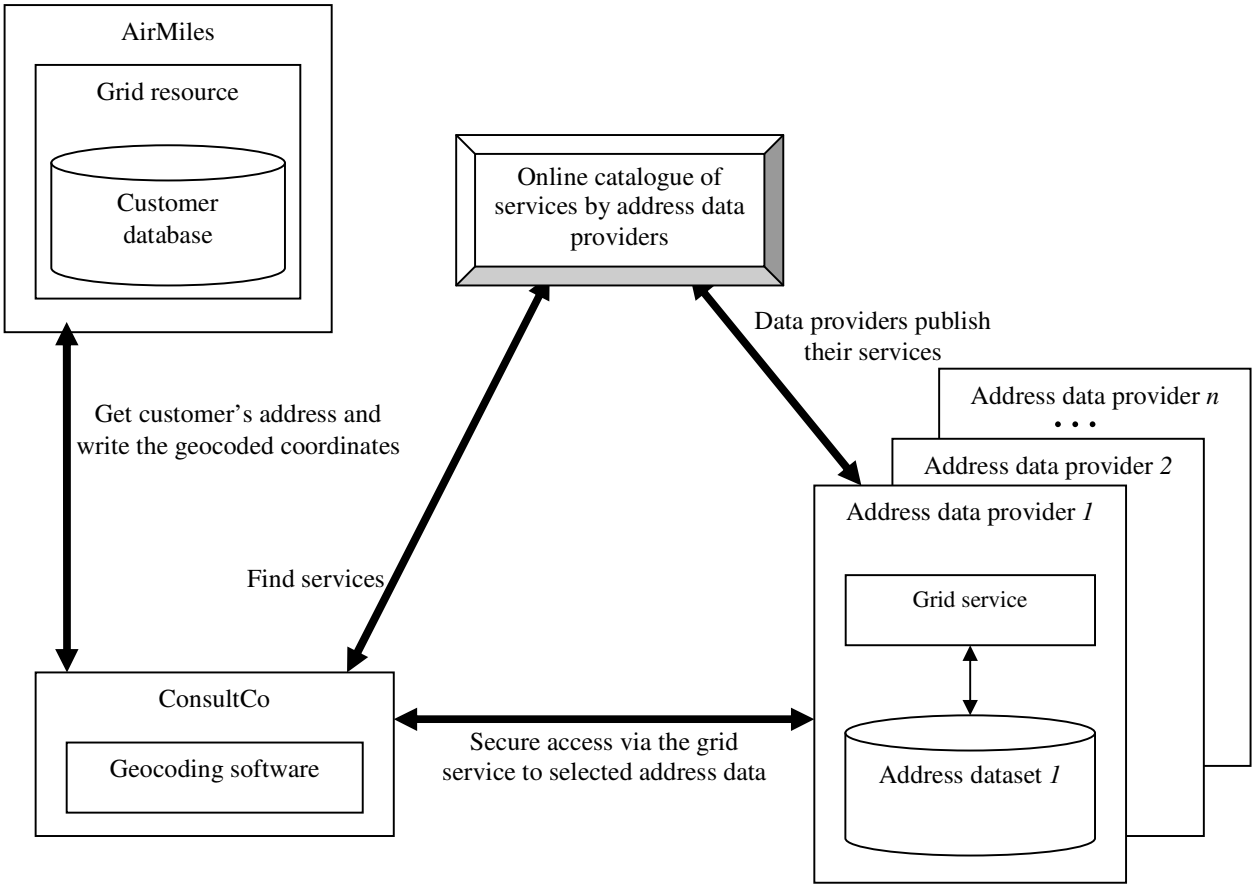


Figure 3. Scenario 2 – Geocoding a customer database

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.

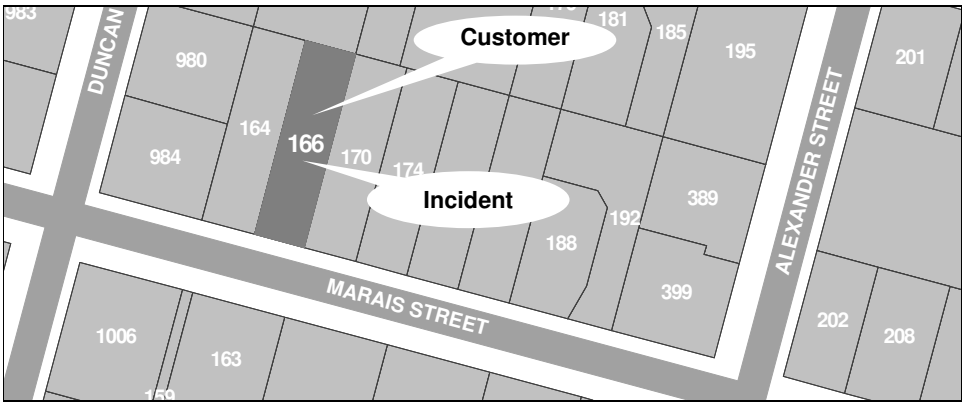


Figure 4. Address data

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.

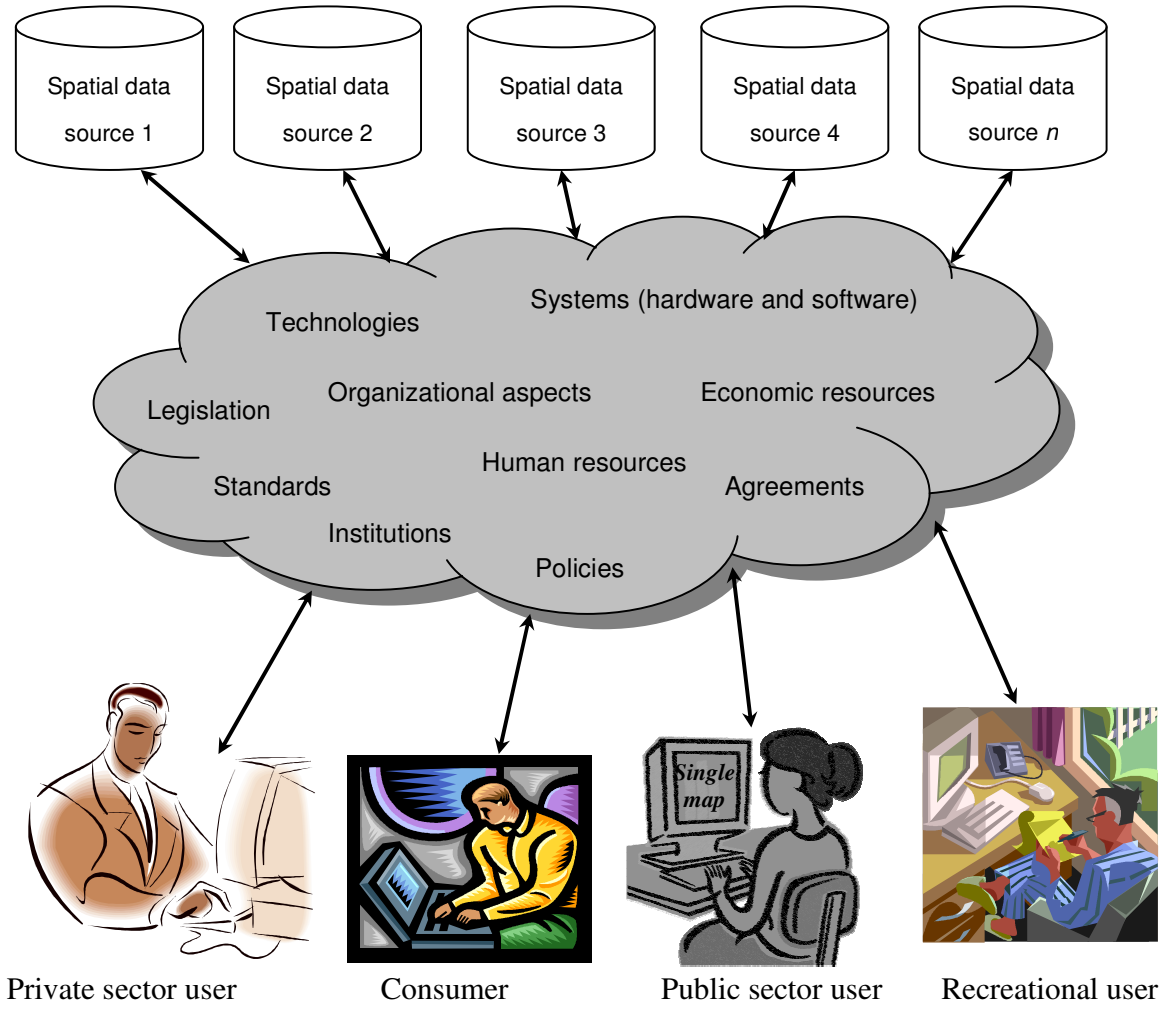


Figure 5. An SDI aims to make multi-source spatial data usable by people

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.

Chapter 3 Data grids

3.1 Introduction

In the first chapter the reader was introduced to data grids. In this chapter more information about grid computing and data grids is presented with the ultimate 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. In reference to Figure 1, this chapter relates mostly to the Computer Science discipline and provides an interpretation of work on data grids in relation to both the Compartimos reference model and, as well as a data grid approach to address databases in national SDI.

In the first section, 3.2, of this chapter the origins, vision and current reality of grid computing are related and concluded with the author's interpretation of an outlook to what the future may hold. Next, in section 3.3, the layered and service-oriented aspects of the common Grid architecture are discussed and related to Compartimos. Subsequently, in section 3.4, data grids, a special case of grid computing, are introduced. A number of existing data grid implementations are analyzed and compared to the requirements for an address data grid in an SDI environment for which the Compartimos reference model, presented in Chapter 4, has been designed. The chapter is concluded in section 3.6 with a discussion of research that is related to the work in this dissertation, confirming that the data grid approach in Compartimos, a reference model for a data grid as enabling platform for address data in an SDI environment, is innovative and new, but also extremely relevant at the current point in time.

3.2 Grid computing

3.2.1 Origins

Grid computing started in the late 1990s as a distributed infrastructure with the main aim of solving specific Grand Challenge applications through high performance computing. The term "Grid" was coined as an analogy to an electrical power grid, envisioning users tapping into a computational grid, similar to consumers currently tapping into an electrical power grid. Since those initial days, the concept of a grid has evolved to address the general need for flexible, secure, coordinated resource sharing among collections of individuals, institutions and resources (Foster and Kesselman 1999).

There is an abundance of definitions for a grid, but one commonly cited definition, and the one that is used in this dissertation, 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.

In other words, a grid integrates and coordinates resources and users that live within different control domains such as different administrative units of the same company or different companies altogether, and addresses the issues of security, policy, payment, membership, and so forth that arise in these settings. A grid is built from multi-purpose, standard, open protocols and interfaces that address such fundamental issues as authentication, authorization, resource discovery, and resource access. A grid allows its constituent resources to be used in a coordinated fashion to deliver various qualities of service, resulting in a combined system that is significantly greater than that of the sum of its parts.

Grid systems are used by virtual organizations (VOs) comprising a set of individuals and/or institutions, having direct access to computers, software, data, and other resources for collaborative problem solving or other purposes. The real and specific problem that underlies the Grid concept is this *coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations*, originating from an emerging need for collaborative problem-solving and resource brokering strategies in industry, science and engineering (Foster *et al.* 2001).

3.2.2 The vision

Foster and his co-authors' vision of how virtual organizations can be enabled to collaborate and share resources in order to achieve a common goal is described in terms of a Grid architecture in the two papers, *The Anatomy of the Grid* (Foster *et al.* 2001) and *The Physiology of the Grid* (Foster *et al.* 2002). This Grid architecture has subsequently evolved into the Open Grid Services Architecture (OGSA) published by the Open Grid Forum (2006), a vision of a broadly applicable and adopted framework for integration, virtualization, and management of resources and services within distributed, heterogeneous, dynamic virtual organizations, which is reflected in the definition of a Grid in the OGSA Glossary of Terms (OGF 2007c):

> A ***system***
> that is concerned with the
> ***integration, virtualization, and management of services and resources***
> in a ***distributed, heterogeneous environment***
> that ***supports virtual organizations*** (collections of users and resources) across traditional administrative and organizational domains (real organizations).

Standardization is a key requirement for realizing this vision of the Grid so that resources that are provided by different vendors and operated by different organizations can be discovered, coordinated and managed in a grid. The Open Grid Forum (OGF) is an open community committed to driving the rapid evolution and adoption of applied distributed computing and the work of OGF is carried out through community-initiated working groups, which develop standards and specifications in cooperation with other leading standards organizations, software vendors, and users (www.ogf.org).

OGSA is a service-oriented architecture that addresses the needs for standardization by describing the requirements and scope of core capabilities that are required to support Grid applications in industry, engineering and science. Communication in the OGSA architecture happens through Grid services, i.e. Web services that provide a set of well-defined interfaces and follow specific conventions. Grid services are specialized Web services and form the Grid-enabling layer between applications and the resources on a lower level, as illustrated in Figure 7. OGSA thus extends the power of the Web services framework, and integrates the Grid and Web technologies to the extent that the distinction between the two is blurring. OGSA is the 'blueprint' for standards-based grid computing (OGF 2008a) and although these Web services are still in the process of being standardized, the Globus Toolkit, an open source software toolkit with implementations of these Web services is fast becoming the *de facto* standard. Compartimos, the reference model that is presented in Chapter 4, is based on the OGSA architecture.

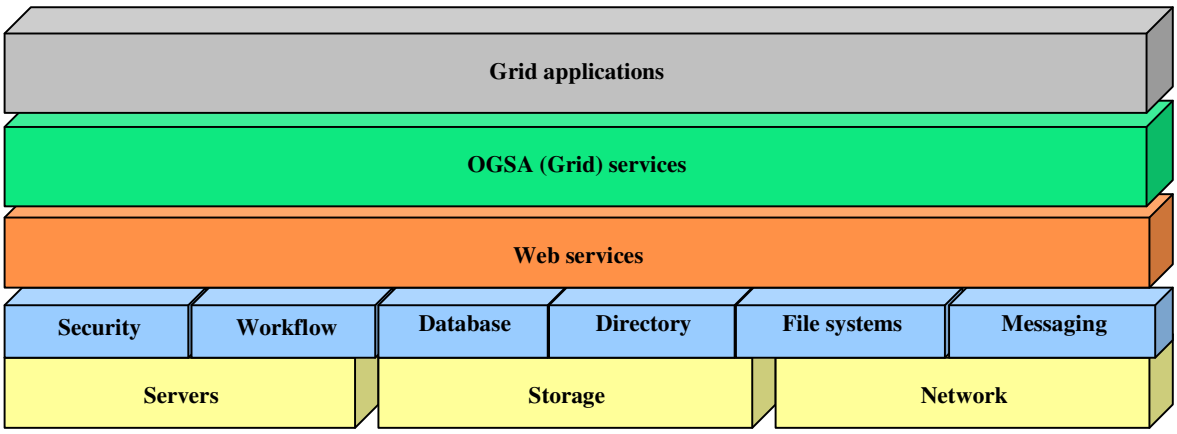


Figure 7. OGSA, adapted from Baker *et al.* (2005)

3.2.3 The current reality

The evolution of a new technology can be divided into an initial developmental phase and a later phase of mass adoption (although some technologies never reach this phase). In the developmental phase the primary concern is the technology itself – how it is built and how it works – and during this stage the users are the experts themselves, the technology is highly specialized and costly to implement. Over time as the technology grows in popularity it is standardized and costs decline until it is adopted by the masses. At this stage the application of the technology, along with the ease of use, reliability, availability and cost become the primary concerns (Wladawsky-Berger 1998).

Grid technology, although popular and fast spreading, still has to become a commodity technology: setting up and maintaining a Grid environment is still quite a complex task. Consequently, skills and resources are highly specialized and therefore limited and expensive. Easy-to-use commodity software is only now starting to emerge with developments such as XtreemOS, an operating system that supports grid applications with native support for setting up and managing virtual organizations, thereby shielding a user from the low-level details of grid middleware and overcoming many of the barriers of entry for establishing virtual organizations. Cloud computing or Web operating systems where users work with Web-based, rather than local, storage and software eliminate the need for skilled resources to set-up a grid in an organization, rather the skilled resources are provided by the 'cloud service provider'. The author suspects this contributes to the current appeal of clouds. Cloud computing is at this point in time becoming more interesting, relevant, and, vendors hope, commercially viable (Coppola *et al.* 2008, Lawton 2008). Commercial viability is a precursor to becoming commodity software.

Without going into a detailed comparison between grids and clouds, with all the recent hype surrounding cloud computing, it is worth noting that clouds have different shapes (Weiss 2007), some of them showing strong resemblance to grids. According to Weiss (2007) the cloud can be seen as a *data center*: cheap, commodity hardware in large numbers controlled by an operating system that is designed to manage resources—hard drive space, memory–to replicate the kinds of intra-server channels that now coordinate events within a single physical machine. This coordination of resources relates to the definition of a grid by both Foster (2002) and later the OGF (2007c). The cloud could be a data center at a single physical location or dozens, hundreds, or thousands of data centers spread around the world, its speed and efficiency is limited by how intelligently it delegates responsibility. Thus alluding to the 'distributed, heterogeneous environment' in the OGF (2007c) definition of a grid and the lack of centralized control in Foster's (2002) definition. Another shape of the cloud is that of the *utility grid*: assuming that a Web application that is hosted in a cloud has been designed intelligently, additional machine instances can be launched on demand so that the

application dynamically, and gracefully, scales up. This is exactly the early vision of Foster and Kesselman (1999) that sees users tapping into a computational grid, similar to consumers tapping into an electrical power grid. Finally, the cloud can provide *software as a service* where processing power is centralized in the cloud and liberates users to choose efficient, uncomplicated access machines that run ultra-thin clients. However, this shape of the cloud centralizes computing power (even it draws on distributed computing resources), whereas a grid aims to coordinate distributed computing resources that are not subject to centralized control.

Data grids based on standard, open, general-purpose protocols and interfaces are still in their infancy because standards and easy-to-use tools are still being developed. Standards development is a slow process, and general adoption and implementation of the standards will take another few years. In the mean time, it is worthwhile to prepare applications and application domains for the world of the Grid, because once Grid standards are in general use and easy-to-use tools are available, Grid technology holds the promise of revolutionizing the world in a fashion similar to the Internet and the Web.

3.2.4 The future

Figure 8 shows the author's interpretation of how the development of software has evolved from tightly coupled software with no software re-use to today's grids for virtual organizations that can be built by dynamically integrating loosely coupled objects on different platforms, in distributed geographic locations and over different administrative domains that deliver services that are beyond the capabilities of an individual organization. Standard, open Grid protocols hold the promise of allowing virtualization of resources until the equivalent of a virtual operating system emerges so that the entire aggregation of heterogeneous architectures can be managed in an automated fashion (Wladawsky-Berger 1998). In fact, Web operating systems, such as Fearsome Engine's ZimDesk (www.zimdesk.com) and Sun Microsystem's Secure Global Desktop (SGD) (www.sun.com/software/products/sgd/index.jsp) are already proof that such virtual operating systems, while not completely open in the sense that any distributed resource can be added and virtualized, are technically and commercially viable (Lawton 2008). Other examples are the Amazon Elastic Compute Cloud (EC2) (http://aws.amazon.com/ec2/) and the Windows Azure operating system Microsoft, announced in October 2008 (http://www.microsoft.com/azure/whatisazure.mspx).

Virtual organizations (VOs) have the potential to change dramatically the way we use computers, much as the Web has changed how we exchange information. Standardized Internet protocols made the Web possible, and standard open Grid protocols hold the promise of fostering unprecedented integration of technologies, applications, files, data, and just about any other IT resource. One cannot predict the future but the author regards it as entirely possible that Grid

technology and Web services will become fully compatible, and that the distinction between the two will eventually fade. Evidence of this fading can be found in the OGSA Glossary of Terms (OGF 2007c): the term 'Grid service' has been deprecated and it is recommended that one refers to it as a 'Web service that is designed to operate in a Grid environment, and meets the requirements of the Grid(s) in which it participates'.

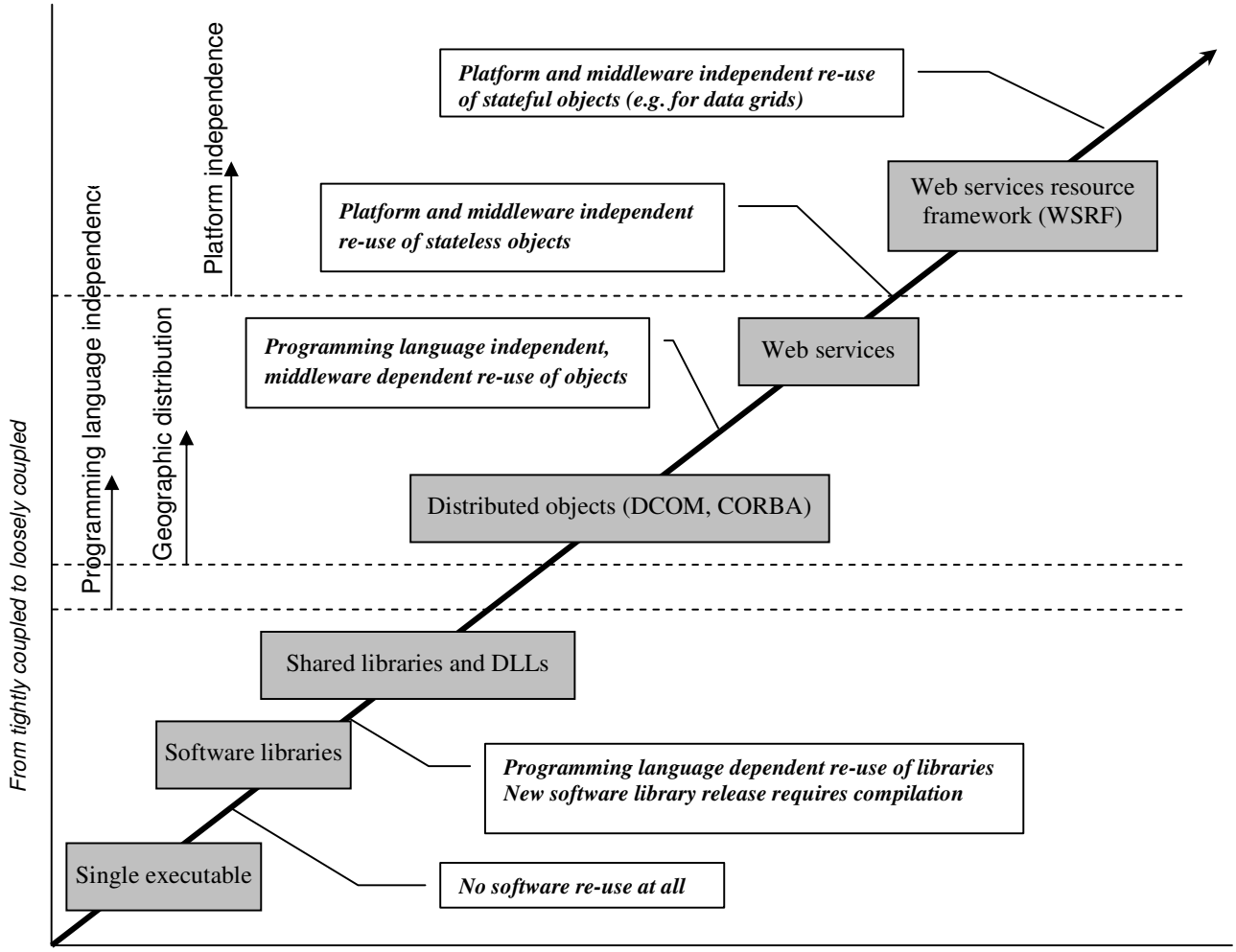


Figure 8. The software evolution

The introduction of the term 'cloud computing' in 2007 for the subset of grid computing that includes utility computing already displays this trend of using Grid technology without being aware of it, a sure sign of advancing from the development phase to mass adoption of a new technology.

The future of the 'cloud' is still unclear and it could pan out to include any or all of the following: word processing and similar applications available as services on the Web; enterprise computing in the cloud where, for example, customer relationship management is supplied as a software service; or the cloudy infrastructures rented out by the likes of Amazon Web Services or Google (Hayes 2008). There is still some debate on what the cloud is and what it's not. To some, the cloud looks like Web-based applications, a revival of the thin-client. To others, the cloud looks like utility computing, a grid that charges metered rates for processing time. Then again, the cloud could be distributed or parallel computing, designed to scale complex processes for improved efficiency (Weiss 2007).

Whatever the relationship between the grid and the cloud, and whatever becomes of the cloud in future, the cloud has certainly gained from past grid research, and with everyone from IBM to Google to Amazon to Microsoft to Oracle to OGC claiming their stake in the cloud through press releases and announcements (IBM 2008, Google 2008, Amazon 2008, Microsoft 2008, Oracle 2008, OGC 2008), it is clear that the cloud, and thus also grid computing, is moving into the everyday realm.

Cloud computing also holds advantages for SDIs in future: the cloud as *data center* could be the platform on which data is shared; the cloud as *utility grid* could provide the computing power that is needed for complex spatial analysis; and the cloud that provides *software as a service* would ensure even wider access to spatial data on any number of thin client devices. While the Compartimos reference model has been designed as a data grid, some Compartimos components, such as the AddressService, the AddressDataAccessService and the VirtualAddressDataService described in Chapter 4, could be deployed in a cloud, probably with some minor modifications.

3.2.5 Note on 'grid', 'Grid' and 'GRID'

In the literature different spellings are observed: 'grid' or 'data grid' (Chervenak *et al.* 2000, Zaslavsky *et al.* 2004, Grimshaw and Natrajan 2005); 'Grid' or 'Data Grid' (Foster *et al.* 2001, Foster 2002, Chervenak *et al.* 2005, Baker *et al.* 2005, OGF 2007c; Venugopal *et al.* 2006); and even 'GRID' (Rajabifard 2005). In this dissertation the term is written with an uppercase, as in 'Grid', when writing about the concept of an infrastructure for resource sharing, while using the lower case 'grid' and 'data grid' for localized or specific Grid implementations. This approach can be observed in the literature but all authors do not consequently apply it.

While the term 'Grid service' has been deprecated in the OGSA Glossary of Terms (OGF 2007c), in this dissertation the term is used to denote a web service in a grid environment.

3.3 The Grid architecture

In this section two aspects of the Grid architecture that are important for Compartimos are presented. *The Anatomy of the Grid* (Foster *et al.* 2001) and *The Physiology of the Grid* (Foster *et al.* 2002), which evolved into the Open Grid Services Architecture (OGSA) from the Open Grid Forum (2006), describe the high level architecture of the Grid. The two aspects of the architecture that are discussed here are the Grid components organized as a layered architecture and the Grid as a service-oriented architecture (SOA). The layered architecture is of interest because in Figure 36 of Chapter 4 each Compartimos component is assigned to one of these layers, thus showing the level of abstraction at which each Compartimos component operates. The SOA aspect of a Grid is of interest since Web service implementation specifications for spatial data discovery and access exist, and these are described and related to Compartimos in the section on Technology choices in Chapter 5 .

3.3.1 The Grid as a layered architecture

The Grid can be described in terms of a number of layers, each at a different level of abstraction, ranging from the fabric layer (the actual hardware) at the lowest level to the application layer (where applications operate in a virtual organization environment) at the highest level. Each layer provides services to the layer above it, and makes use of services that are provided by the layer below it. Each layer also provides a virtualization of the resources on the lower level, e.g. the differences between hard disks from different vendors are accommodated by the operating systems in the Grid fabric layer, and on the application layer a storage resource or computing resource is requested, regardless of all the intricate details of the actual device, the discovery mechanisms to locate it and the communication protocols to use it.

In Compartimos there is also abstraction and virtualization of the distributed address data sources, as well as services and protocols that determine behavior and coordination on the collective layer, i.e. for the collection of address datasets. The layers in Figure 9 below provide a reference for this allocation of Compartimos components to individual grid layers, which will be discussed in Chapter 4 . For the purposes of describing the layered architecture of the Grid, the layers presented by Venugopal *et al.* (2006) and Foster *et al.* (2001) are merged into four main layers of the Grid architecture, illustrated in Figure 9 below, and described as follows.

Application layer. On this layer domain-specific applications operate within a virtual organization (VO) environment to achieve the VO's collaborative goal, such as climate modeling, hybrid earthquake engineering experiments and physics data analysis; or mapping and geocoding in the case of the work in this dissertation.

Grid resources and services layer. Consists of protocols for the secure negotiation, initiation, monitoring, control, accounting and payment of individual resources (the *Resource* layer according

to Foster *et al.* 2001) as well as collections of resources (the *Collective* layer according to Foster *et al.* 2001). This layer provides the required level of abstraction of individual resources so that the Grid can provide a wide variety of behaviors for collections of resources that are based on these resource abstractions. Examples are resource discovery, resource brokering, job scheduling, replication and replica management, resource monitoring, and accounting services.

Connectivity and communication layer. Consists of the communication and authentication protocols that are required to interact with the resources on the Fabric layer. These protocols are mainly drawn from the TCP/IP protocol stack but are augmented with protocols incorporating specific Grid requirements such as the GridFTP protocol, which provides efficient transfer of large data files in a grid. Existing authentication protocols such as public key infrastructure (PKI) in the form of X.509-certificates are integrated and extended on this layer.

Fabric layer. Consists of the physical resources to which the Grid coordinates shared access. These are computational resources (clusters, supercomputers), storage resources (RAID disks, tape archives), data resources (databases, files) and instruments (sensors, telescope, accelerator). Each of these resources is controlled by software such as the operating system, file system and/or database management system.

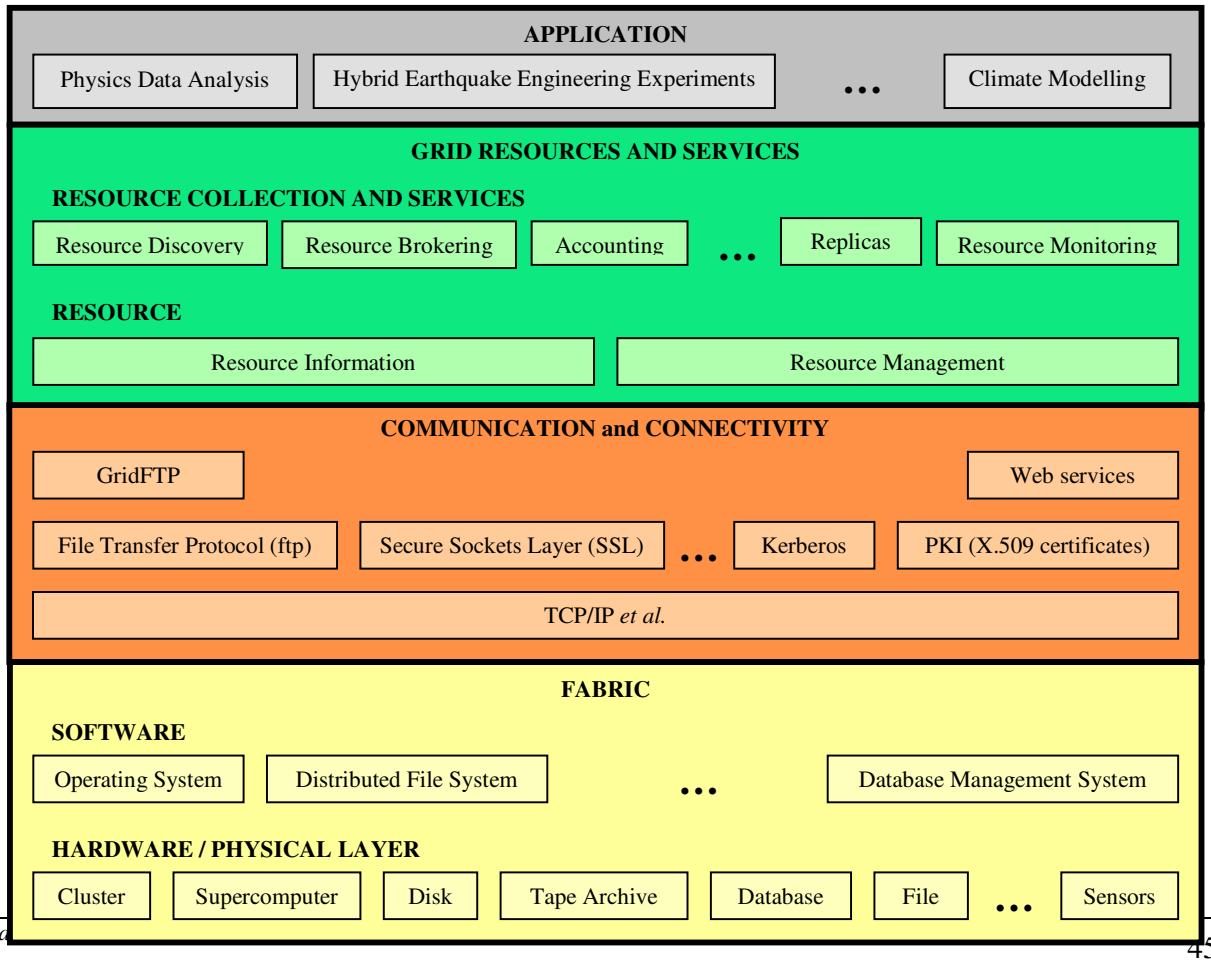


Figure 9. The four main layers of the Grid architecture

3.3.2 The Grid as a service-oriented architecture

A service-oriented architecture (SOA) refers to the specific style of building a reliable distributed system that delivers functionality as services, with the additional emphasis on loose coupling between interacting services (OGF 2007c). An SOA is typically implemented by a set of Web services that provide the capabilities and behaviors of the system. OGSA is a service-oriented architecture in which the core capabilities and behaviors are described as a set of services, the Grid services. These Grid services are loosely coupled peers that, either singly or as a part of an interacting group of services, realize the capabilities and behaviors of OGSA (OGF 2006).

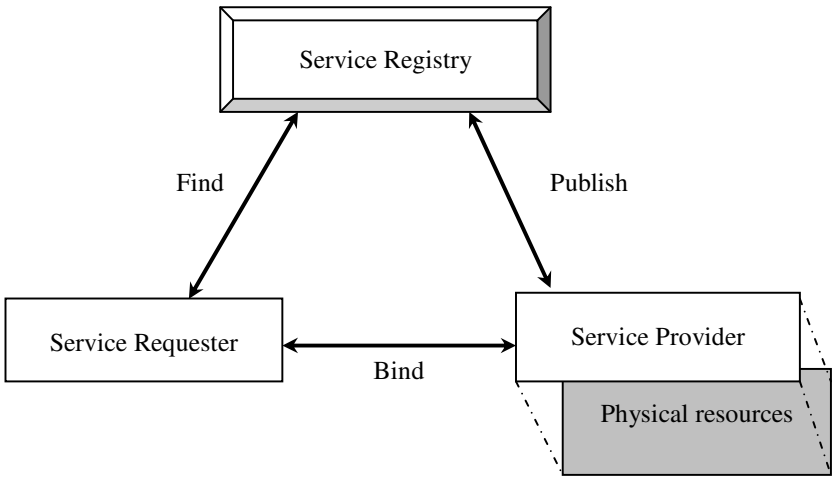


Figure 10. Service-oriented architecture

A service-oriented architecture is further based on the concept that a service provider publishes its services at a service registry. A service requester finds details about a specific service at the registry, and then proceeds to bind to the service at the service provider and starts interacting with the service at the service provider. This concept is illustrated in Figure 10. OGSA is a specific profile of the core Web Service standards that includes the Web Services Description Language (WSDL), Simple Object Access Protocol (SOAP) and Universal Description Discovery and Integration (UDDI) that have been specified by the W3C.

Web Services (WS) standards do not support and were not designed to meet all Grid requirements, but the Grid community is actively involved in the development and evolution of WS

standards to revise, modify and extend existing standards and to introduce new standards where applicable. For example, Web services are stateless, but OGSA requires stateful Web services. The Web Services Resource Framework (WSRF) and WS-Management families of standards, published by OASIS, were prompted by the Grid community (Baker *et al.* 2005). While statefulness is a fundamental requirement for the Grid, other application domains, such as general-purpose Web servers, benefit from these standards as well.

Compartimos follows a service-oriented approach, i.e. there is a registry of discoverable AddressDataAccessServices, published on the address data grid by address data providers, which the VirtualAddressDataService discovers and coordinates in order to achieve the virtual address dataset. These services are presented and discussed in Chapter 4 . The SOA aspect of a Grid is further of interest since OGC Web service implementation specifications for spatial data discovery and access exist, and these are described and related to Compartimos in the section on Technology choices in Chapter 5 .

3.4 Data grids

3.4.1 What is a data grid?

A *data grid* is a special kind of Grid in which data resources are shared and coordinated. In OGSA (OGF 2007c) a *data resource* is defined as an entity (and its associated framework) that provides a data access mechanism or can act as a source or sink of data. These data resources are typically heterogeneous in terms of their syntax and semantics. Examples of data resources are flat files, tables in a relational database, sensors or data streams. Referring back to Foster's (2002) definition of a grid, in a data grid

1. the individual *data resources* that are shared on the grid live in different control domains and consist of flat files, tables in relational databases, or other sources of data;
2. these constituent *data resources* are coordinated to deliver a *data service* that is significantly greater than the sum of its parts and,
3. all of this is achieved through standard, open, general-purpose *data protocols and data interfaces*.

Or, adapting the definition of a grid in the OGSA Glossary of Terms (OGF 2007c):

> A ***system***
> that is concerned with the
> ***integration, virtualization, and management of data services and data resources***
> in a ***distributed, heterogeneous environment***
> that ***supports virtual organizations*** (collections of users and data resources)
> across traditional *administrative and organizational domains (real organizations).*

Data grids are used for the sharing and integration of distributed data that are managed and administered independently, also referred to as *data federation*. Data grids are also applied in areas of science, technology and commerce where there is a need for efficient access to, and the movement and management of, large quantities of data in a distributed environment, also known as *data-intensive environments*.

Compartimos is a reference model for an address data grid in which distributed heterogeneous sources of address data are managed and administered independently of each other. The data grid provides coordinated access to these distributed heterogeneous sources of address data and is therefore an example of a data grid that is used for *data federation*.

3.4.2 OGSA-Data Access and Integration (OGSA-DAI)

OGSA-Data Access and Integration (OGSA-DAI) is a service-oriented architecture for database access over the Grid that allows for the integration of heterogeneous databases into an OGSA-type grid, and this technology is therefore relevant to Compartimos where coordinated access to distributed heterogeneous sources of address data is provided. OGSA-DAI originated from a need to include in Grid implementations data that is stored in a DBMS. Current DBMSs already provide a large range of functionality to securely store, query and maintain large volumes of data, but none of them have been OGSA Grid-enabled. Rather than to build an OGSA Grid-enabled DBMS from scratch, the principle behind OGSA-DAI is to provide the necessary middleware that will 'OGSA Grid-enable' existing DBMSs. As such OGSA-DAI has to reconcile DBMS implementation differences (IBM DB2, Oracle, MS SQLServer, etc.) and accommodate the variety of database paradigms (relational, object, XML, etc.). Additionally, OGSA-DAI provides distributed query functionality (OGSA-DQP) that allows a user to send a single data request to an OGSA Grid-enabled data grid and the OGSA-DQP then takes care of coordinating the request among the different data resources (Antonioletti *et al.* 2005).

Figure 11 shows the layered architecture of OGSA-DAI. The OGSA-DAI Basic Services are implemented as OGSA Grid Data Services and access the underlying databases using drivers and also provide for data formatting, data delivery and request handling. A Grid application can either use the OGSA Grid Data Services (OGSA-GDS) directly or use the OGSA-DQP to coordinate access to the multiple databases. OGSA-DAI implements its services as OGSA Grid Services, and is

therefore also a service-oriented architecture.

OGSA-DAI has been used in a number of data grid implementations, including the eDiaMoND project described in section 3.5. In some ways OGSA-DAI works like Open Database Connectivity (ODBC) and Java Database Connectivity (JDBC): ODBC and JDBC provide a standard programming interface to any DBMS, and as long as a driver exists for a specific DBMS, that DBMS can be accessed through the standard ODBC/JDBC interface. OGSA-DAI provides the standard interface to the various databases (in various DBMSs and paradigms) in a data grid. In Compartimos the *AddressDataAccessService* performs the role of such a 'driver'.

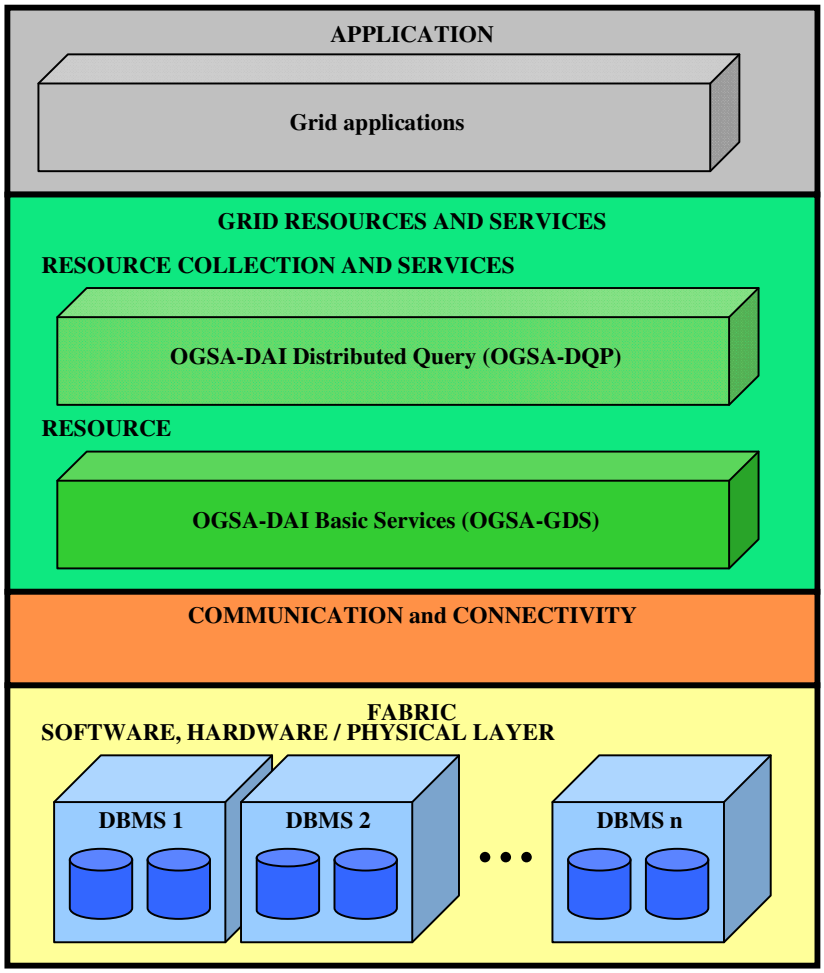


Figure 11. The OGSA-DAI layered architecture, adapted from Antonioletti (2005) to show the four main Grid layers of Figure 9

In an SDI environment address data is stored in data resources that include relational databases (such as Oracle and Microsoft SQL Server), spatial databases (such as Oracle Spatial and ArcSDE),

as well as proprietary geographic files (such as ESRI .SHP and MapInfo .TAB files). In order to integrate the data from the multiple heterogeneous sources of address data described in Chapter 2, both the syntactic (Oracle Spatial, ESRI SHP, etc.), as well as the semantic (address data model) differences have to be accommodated.

The OGC has published a specification for a Web Feature Service (WFS) (OGC 2005) that provides a platform-independent data access interface to features (representations of real world objects) in a spatial dataset. A draft ISO standard, ISO 19142 (draft), *Geographic information – Web Feature Service*, is also available. One potential approach to syntactic interoperability in an address data grid is to make use of a single OGC WFS that is able to interpret different types of address data resources (Oracle Spatial, ArcSDE, etc.), but returns address data in a common format such as the standard open Geography Markup Language (GML), an XML grammar for the modeling, transfer, and storage of geographic information (ISO 19136:2007). An alternative approach is to make use of many OGC WFSs, one per type of spatial data resource, each translating between the proprietary format and a common format such as GML. The same goes for semantic interoperability: either a single service is able to interpret all kinds of address data, meaning that 'understanding' an additional address dataset requires an updated implementation of the service; or there is one service for each address data model, meaning that an additional address dataset requires the registration of an additional service. The latter approach scales better for many different formats and/or models and is therefore followed in the Compartimos reference model. Figure 12 below illustrates the two different approaches.

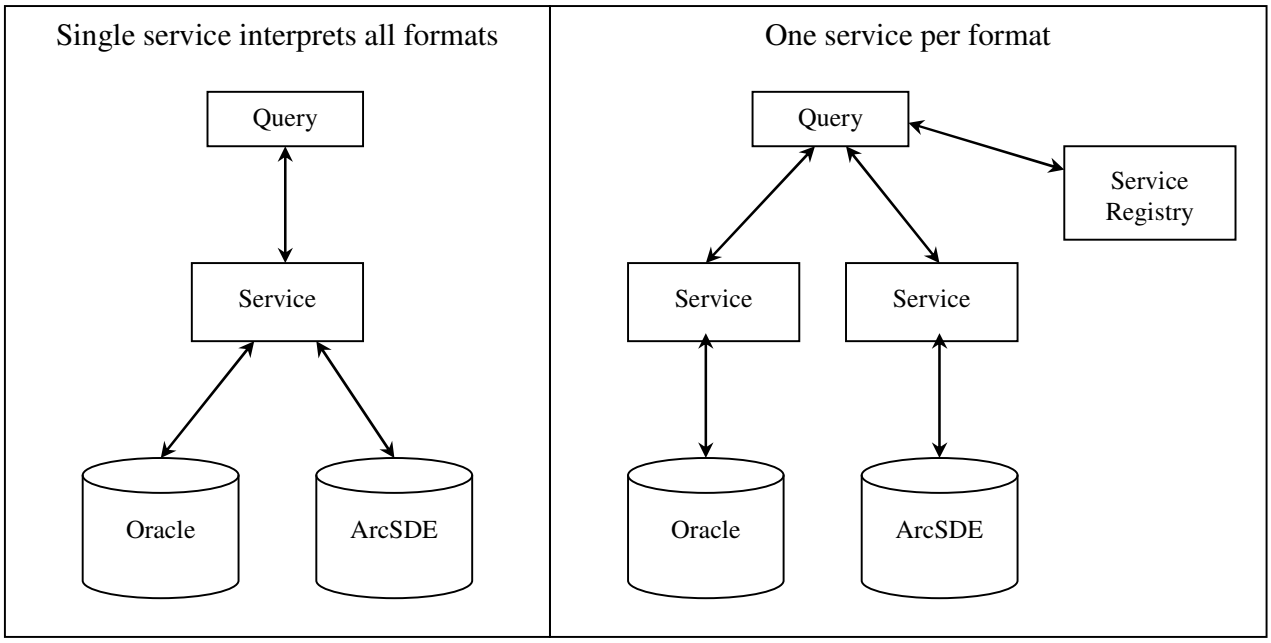


Figure 12. Two approaches to resolving syntactic heterogeneity

3.5 Examples of data grid implementations

In this section the following existing data grid implementations are described:

- the Laser Interferometer Gravitational Wave Observatory (LIGO);
- the Earth System Grid (ESG);
- the e-DiaMoND project; and
- the Geoscience network (GEON).

Table 4. Existing data grid implementations (author's summary)

	LIGO	Earth System Grid (ESG)	e-DiaMoND	GEON
Application domain	Physics and astronomy	Climate modeling	Breast cancer treatment	Earth sciences
Region	United States	United States	United Kingdom	Northern America
Number of data sites	Two	Around 10 centers and laboratories	Scalable to 90+ Breast Care Units (BCUs) in the UK	3 GEON data nodes, 15 GEON points of presence
Total data volume	One terabyte per day, ca. 365 terabytes per year	250 terabytes until 2006, ca. 70 terabytes per year	Estimated 480 terabytes per year, when fully operational	Each data node can store 4 terabytes of data
Metadata	Descriptive metadata about the data in the files (in a relational database)	Climate model metadata (in a relational database)	Patient data and metadata on image files (in a relational database)	Metadata about data made available by providers (in a relational database)
Format of data resources	Files with data from the LIGO detector	Files containing climate research data	Image files with mammography	Relational data, ESRI .SHP files, LiDAR
Size of individual data item	1-100 megabytes per file	Unknown	Estimated 75 megabytes per image file	Varies considerably, depending on what a user uploads
Number of data items	More than 40 million files	Millions of files	1000 cases	Around 4,500 (searching the portal on March 2008)
Interaction	Portal www.ligo.org	Portal www.earthsystemgrid.org	Service registry and Web services	Portal www.geongrid.org, as well as Web service registry
Software	Globus Toolkit, Lightweight Data Replicator (LDR)	Globus Toolkit, OPeNDAP-G (Grid-enabled Open-source Project for a Network Data Access Protocol)	Globus Toolkit, OGSA-DAI, IBM: DB2, Content Manager, Visual Age C++, WebSphere Application Server, Apache TomCat	Globus Toolkit, OGSA-DAI, Storage Resource Broker (SRB), ROCKS, GridSphere, PostgresSQL, IBM DB2, MySQL, ArcIMS, GRASS, ArcSDE

These implementations were selected specifically to illustrate variety in data resource types, data volumes and client interaction in order to provide a broad comparative base for Compartimos. A summary overview (prepared by the author) of some of the characteristics of these implementations is supplied in Table 4. In section 3.6 these existing data grid implementations are related to the work on Compartimos in this dissertation in order to position this work in the bigger context of data grids. In Chapter 4 the summary is repeated, this time including the requirements for an address data grid in an SDI, so as to distinguish the Compartimos requirements from other data grid implementations.

3.5.1 Laser Interferometer Gravitational Wave Observatory (LIGO)

LIGO is a facility dedicated to the detection of cosmic gravitational waves and the harnessing of these waves for scientific research. Gravitational waves are ripples in the fabric of space and time produced by violent events in the distant universe, for example by the collision of two black holes or by the cores of supernova explosions. These ripples in the space-time fabric travel to Earth, bringing with them information about their violent origins and about the nature of gravity. As such LIGO is a scientific tool to assist research in both physics and astronomy (http://www.ligo.caltech.edu/, Chervenak *et al.* 2005).

The LIGO facility includes three interferometers at two sites that generate approximately one terabyte of data every day. Refer to Figure 13 for photos of the facilities. The data generated must be scientifically analyzed for it to be of any value. The analysis is computationally intensive (some classes of astrophysical searches can require hundreds of Teraflops), and the data volume itself is huge (1 TB per day).

LIGO Hanford Observatory
Hanford, Washington, United States

Livingstone Observatory
Livingstone, Louisiana, United States

Figure 13. LIGO installations (http://www.ligo.caltech.edu/)

Nine sites within the LIGO collaboration (each operated independently) currently provide computing facilities based on commodity cluster computing. The scientists who have the expertise to perform this analysis are spread across 41 institutions on several continents, and this community is growing all the time. The key challenge for LIGO is to get the data from the LIGO detectors to the sites where analysis happens and to make those sites accessible to the participating scientists.

The data management challenge faced by LIGO is therefore to replicate approximately one TB/day of data to multiple sites securely, efficiently, robustly, and automatically; to keep track of

where replicas have been made for each piece of the data; and to use the data in a multitude of independent analysis runs. The nine sites each use mass storage systems, but different systems are used at different sites. Scientists and analysts need a coherent mechanism to learn which data items are currently available, where they are, and how to access them.

The work on the Lightweight Data Replicator (LDR) was developed for the LIGO system and the knowledge gained from its development served as input to the Data Replication Service (DRS) of the Globus Toolkit, which provides a pull-based replication capability similar to that provided in the LIGO LDR system. (www.globus.org).

3.5.2 Earth System Grid (ESG)

ESG is a data grid that connects important U.S. repositories of climate model data that are geographically distributed at a number of national laboratories and research centers. In 2003 alone, climate change research sponsored by the US Department of Energy (DOE) has produced at least 72 terabytes of scientific data that is stored across several of the DOE sites. The goal of the ESG is to improve the daily management and tracking of this data by facilitating data publishing by climate modelers, and by facilitating access to the data by the worldwide climate community.

The main entry point to the ESG is through a Web portal on which users search or browse ESG catalogues to locate desired datasets, with the option of browsing both metadata about experiments as well as individual files and their contents. Users select datasets or individual files to be retrieved or request an “aggregation” – a specific set of variables subject to a spatiotemporal constraint. Selected data can be downloaded to the user’s system for analysis.

The ESG portal provides a central location for enforcing authentication, authorization and accounting (AAA) services, and brokers the formulation and submission of user data requests (transfer, download, sub-setting) among the distributed data nodes. The portal also provides the interface through which authorized data providers publish datasets into the system. Data are stored either on disk farms for faster high-performance online access to frequently requested datasets, or on deep archives for less frequently accessed datasets (Foster *et al.* 2006). Refer to Figure 14 for the ESG topology.

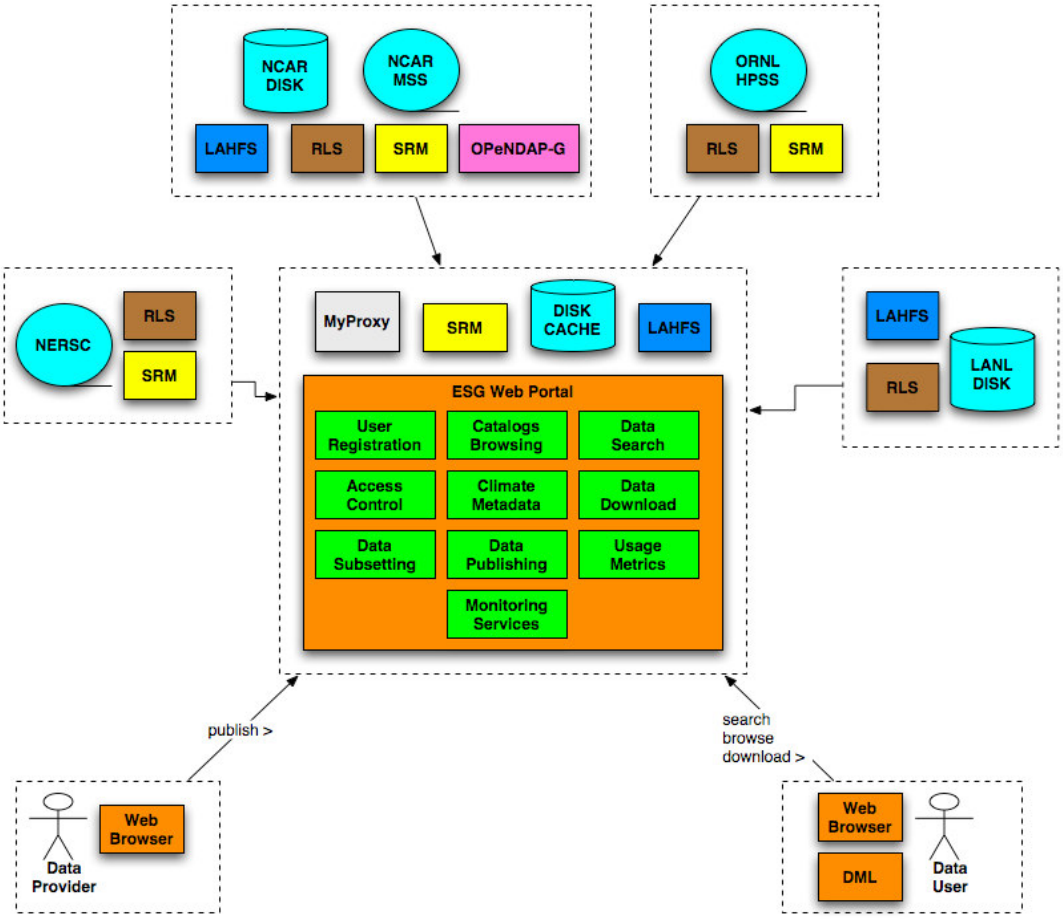


Figure 14. The ESG topology (Foster *et al.* 2006)

3.5.3 e-DiaMoND

e-DiaMoND is a British research project that developed a prototype Grid infrastructure for mammogram databases that are stored at geographically distributed independently managed breast care units (BCUs). The goal with the infrastructure was to enable the sharing of mammography archives, the easy, dynamic and real-time movement of information with high levels of security to facilitate the sharing of workloads, and collaboration of radiologists without being in the same physical location. Figure 15 shows examples of mammography.

The core e-DiaMoND system consists of middleware and a virtual image store comprising physical databases, each owned and administered by a different organization (the BCUs). The e-DiaMoND grid is formed by participating BCUs coming together as a virtual organization to unite their individual databases as a single logical resource, the virtual image store. Clients interact with the e-DiaMoND registry to discover services, and then interact with the grid through these services, which can be divided into data services—that allow each hospital to see all of the data owned by the

participating nodes—and compute services—that can perform potentially complex and long-running calculations on the image data.

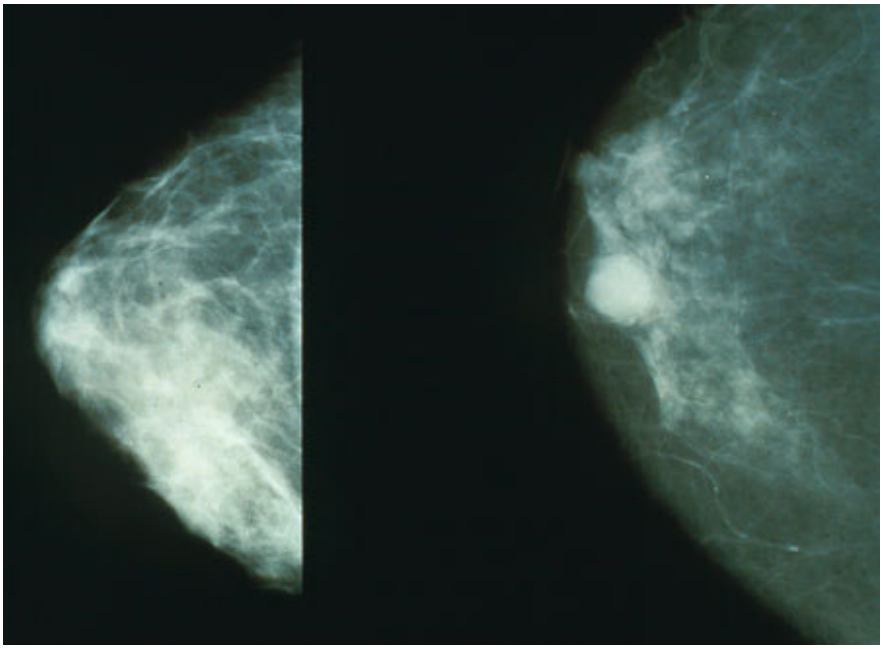


Figure 15. Example mammography image: normal (left) versus cancerous (right)

3.5.4 GEON

The Geoscience network (GEON) project aims to develop cyber-infrastructure in support of integrated research to gain a more quantitative understanding of the 4-D evolution of the North American lithosphere (the crust and uppermost mantle of the Earth). Scientists in this area are providers (computing, storage, etc.) as well as consumers (e.g. for analyses) of resources and are themselves distributed at a number of organizations. The themes that provided the initial guidelines for realizing this cyber-infrastructure illustrate the need to accommodate a very wide range of data variety (www.geongrid.org):

(1) gravity modeling of 3-D geological features such as plutons, using semantic integration of (igneous) rock and gravity databases, and other geological and geophysical data;

(2) study of active tectonics via integration of LiDAR data sets, data on distribution of faults and earthquakes, and geodynamics models; and

(3) study of lithospheric structure and properties across diverse tectonic environments via the integration of geophysical, petrologic, geochronologic, and structural data and models.

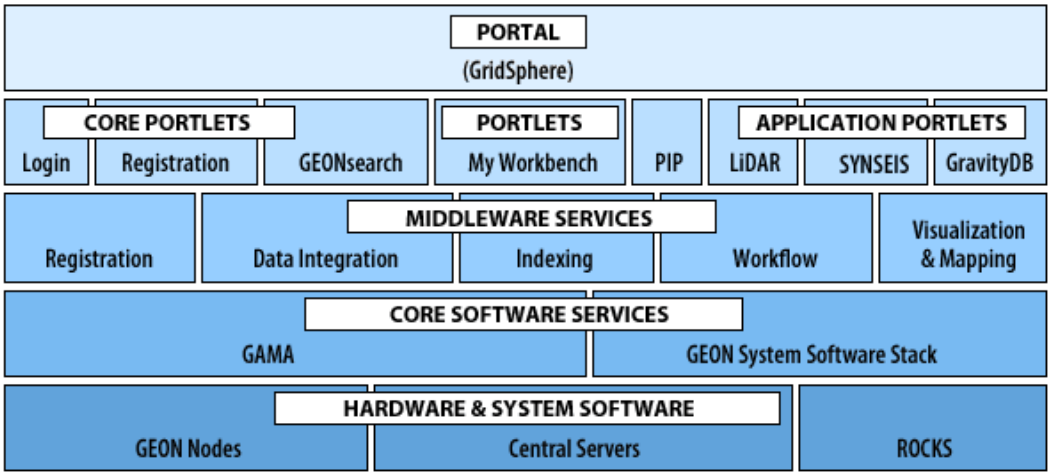


Figure 16. The various components of the GEON system

The GEON architecture consists of a single GEON portal and a number of distributed GEON sites and GEON service providers. The portal provides a single point of entry into the distributed GEON environment. Refer to Figure 16. GEON is based on a service-oriented architecture (SOA) in which remote sites and service providers communicate via Web services. All servers, at the central site as well as remote sites, run the same software stack, which is configured and deployed using ROCKS, a server clustering toolkit.

A site deploys a Point of Presence (PoP) node to host data, tools, and other resources contributed into the GEON network. Alternatively data, tools and other resources are deployed ("hosted") on the central site and hosted from the central location. Resources (data and tools) may also be contributed into the GEON network by providing Web service interfaces to the same. In this case, there is no requirement to install a PoP node at the remote site. The provider only needs to ensure that their Web services are able to interoperate with the GEON SOA environment.

3.6 Related work

3.6.1 Compartimos in relation to the examples of data grid implementations

In this section the work on Compartimos is compared to the examples of data grid implementations that were described in the previous section. Compartimos itself is presented in detail in Chapter 4 and this section serves to relate work on Compartimos to other existing data grid implementations. This related work falls mainly into the Computer Science discipline and only some of it overlaps with GISc. Other related work in the GISc discipline specifically was discussed in Chapter 2.

Compartimos, a reference model for an address data grid in an SDI, falls into the *application domain* of Geographic Information Science (GISc) but it is not in the scientific or research domain; rather it attempts to solve a real-world problem. ESG and GEON focus on data relating to the physical aspects of the Earth, while addresses are assigned by people according to certain rules and conventions, often in relation to (invisible) administrative boundaries and man-made features. Address data is influenced by the authorities that define these boundaries, as well as construct and maintain address-related infrastructure such as streets, houses and buildings.

The *regions* listed for the data grid implementations in Table 4 in the previous section span a single country or continent, but could be expanded globally should the need arise. The need for address data is firstly on a national, and later on an international scale. Consider the European situation where international seamlessly integrated address data for the whole of Europe is required in terms of the INSPIRE Directive. For the Compartimos implementation sample data of different address types (street addresses, site addresses, etc.) from multiple local sources in South Africa was used and the resulting consolidated dataset spans a single country.

Many of the early Grid implementations, including the data grid implementations described in the previous section, were built with the *purpose* to fulfill the demand for high performance computing and data-intensive applications in scientific research. This is a natural development since experts initiating the Grid computing concept were centered at universities and research centers and were probably in a better position to understand the scientific requirements. However, the initial concept of the Grid has expanded to the wider area of resource sharing among members of virtual organizations and is not restricted to scientific applications anymore. The OGF has passed some of the components that were identified for standardization in OGSA onto the W3C and OASIS (an industry consortium), from which one can conclude that Grid computing is progressing on its way to reach a wider audience. Refer also to current hype about cloud computing, discussed earlier in section 3.2.3. The main *purpose* of Compartimos is not high performance computing, but data integration and federation, although high performance computing could be a side benefit from the grid.

The *number of data sites* in the data grid implementations that are described in the previous section are limited, ranging from two to ninety. Address data typically resides at the local authorities in a country, the number of which in a country differs, but is very often more than a hundred. South Africa, for example, has around 260; in the UK there are 376 in England and Wales, and 32 in Scotland (Coetzee *et al.* 2008b); Australia has 750 (Jacoby 2002); and Denmark had 271 that were recently rationalized to 98 (Lind 2004). In this regard Compartimos is more like a sensor grid that has hundreds or thousands of sources of data, the sensors. However, in a sensor grid new data is continuously generated, while in Compartimos each local authority shares a comparatively stable

database of addresses.

The *total data volumes* of the LIGO, ESG and eDiaMoND data grids suggest that they are data intensive, whereas the volumes of address data are usually smaller: even the conservative estimate of assuming that there is a spatial address for every person on earth, each taking up 5-10K bytes (extremely conservative estimate: a typical address record will take up even less space) of storage space results in (only) a total of 32TB to 66TB of data. While this data might change over time, there is not a continuous stream of new address data. Thus comparatively, address data volumes are small.

The address data grid stores *metadata* about data providers and the content of their data. This is stored in a relational database and is available through the *CatalogueService*. Metadata for each individual address record is stored at the data provider. In this regard Compartimos is similar to the GEON grid but differs from the LIGO and eDiaMoND data grids. In Compartimos the design of the data model for the metadata is based on the ISO 19115:2003*, Geographic information – Metadata* standard.

Although the data in the ESG is linked to locations on the Earth, it is limited to climate research data, which is often better represented by raster-type data. The spatial simplicity of an address feature – very often represented as a point feature – on the other hand is best represented by vector data. The *format of data resources* in Compartimos compares well to the GEON grid where data is stored in proprietary GIS files, such as ESRI .SHP files, of which the format is similar to the type of address data that one finds at a local authority. Both ESG and GEON store and provide geographic data, but in relation to Compartimos, it is a different kind of geographic data. ESG and GEON focus on data relating to the physical aspects of the Earth, while address data in Compartimos is assigned by people according to certain rules and conventions, often in relation to administrative boundaries and man-made features. Thus address data is influenced by the authorities that define these boundaries and construct and maintain address-related infrastructure such as streets, houses and buildings.

The size of an address record is in the range of 5-10K, which is a lot smaller than the *size of individual data items* in the data grids described in the previous section: 75MB and 100MB data files in the LIGO and eDiaMoND grids respectively.

Interaction with Compartimos is through a portal, similar to the LIGO, ESG and GEON grids, but Grid services can also be registered in a service registry from where third party organizations can discover them in order to use them, similar to the GEON and eDiaMoND data grid. Thus Compartimos has a service-oriented architecture.

3.6.2 Other related work

In this section data grid related work that is similar to Compartimos is described. The work confirms that the research described in this dissertation is relevant and novel.

Zaslavsky *et al.* (2004) describe *Smart Atlas*, a GIS-based atlas environment enabling users to discover, access, visualize and query heterogeneous brain images and image markup. These brain images are organized into atlases of spatial data with 2-D and 3-D visualization techniques. The spatial data sources for Smart Atlas are web-enabled and include ArcIMS feature and image services and distributed Grid sources. The Smart Atlas client can be invoked from the Biomedical Informatics Research Network (BIRN) portal at www.nbirn.net. Even though Smart Atlas does not represent geographic data as such, it is relevant to Compartimos from the point of view that it includes ArcIMS data sources, which are a common format for storing address data.

The *multi-node and multi-data source grid GIS (MMG^{2}IS)* is a Grid system that was developed by integrating Grid technology with networked GIS systems (Wang *et al.* 2004). Grid technology that was used includes the Globus Toolkit and Grid development tools such as the Java Commodity Grid Kit and the IBM Grid application framework for Java. MMG^{2}IS employs virtualization through a system architecture that is based on a virtual machine: the different underlying machines are virtualized into a single virtual machine representing a single Grid-enabled operating system. Compartimos makes use of the concept of virtualization of address data sources, albeit at different levels and for different purposes. The MMG^{2}IS is further relevant because it integrates spatial data. It was, however, developed for a data-intensive environment whereas the purpose of Compartimos is data integration and federation.

Zhao *et al.* (2004) present a geospatial registry approach in which the OGC WRS (Web Registry Service), a de facto standard that supports the publishing of, and run-time access to, geospatial resources, as a wrapper, is used to extend the capabilities of the conventional Grid Metadata Catalogue Service (MCS) to the processing of geospatial queries against multiple heterogeneous spatial data sources and services. The implementation of this approach was used in the NASA Grid Data Service environment. This report is relevant to the Compartimos catalogue service.

The Distributed Earth Observation System Information Service (DEOSIS) (Aloisio *et al.* 2005b) was developed by the University of Lecce and aims at managing and accessing earth observation and geospatial heterogeneous data sources in a Grid environment. A Grid-based architecture provides a secure, scalable and pervasive environment for earth observation and geospatial data management among several virtual organizations. In this grid the metadata model is based on ISO 19115:2003, *Geographic information – Metadata,* which is interesting because the Compartimos catalogue data

model is also based on ISO 19115.

The geographic data for which the DEOSIS data grid was developed, is mostly Earth observation data (similar to the ESG and GEON grids), which differs from address data: Earth observation data is often gathered with an 'eye-in-the-sky', be it aerial photography or satellite imagery. Addresses on the other hand have a close link to land administration and legal processes, and are therefore usually assigned by people and maintained by the relevant authorities.

Xue *et al.* (2008) write about the remote sensing information grid node (RSIN), which is a tool for dealing with climate change and quantitative environmental monitoring. In their case data-intensive retrieval of remote sensing information is required. They have developed a failure management strategy using a throughput-estimated model for data nodes. The failure management strategy is of general interest but this report is of specific interest to the work in this dissertation because spatial data is used in a grid but the application is for data intensive operations, and not for data federation as in the case of Compartimos.

Hua *et al.* (2005) present a design and small-scale implementation of a spatial data grid for a few hundred MBs as a proof of concept that grid computing technology can be applied to share distributed spatial data. They propose their spatial data grid as the GIS infrastructure for sharing spatial data comprehensively and thereby eliminating information islands. In this regard, their research supports the proposed data grid approach to sharing address data in an SDI, discussed in Chapter 6. In conclusion, Hua *et al.* recommend that more research should be done on the combination of grid computing with GIS. Compartimos is an example of such multi-disciplinary research that aims to apply grid computing to the sharing of distributed spatial data.

Ghimire *et al.* (2005) propose the integration of geographic information Web services with mobile agents and the grid in order to deliver large size spatial datasets over limited bandwidths. They recommend that the GI community should reconsider its adherence to the HTTP POST and GET communication patterns and move towards a SOAP based communication, which is a necessary and urgent step that would pave the road for a number of important performance issues, and from the point of view of Compartimos it would also enable seamless integration into the OGSA architecture. Aydin et al. (2008) report on an OGC WFS implementation that supports the three mandatory operations of the WFS implementation specification through a WSDL interface. Work on this recommendation is also now in progress at OGC, as reported in the *Summary of the OGC Web Services, Phase 5 (OWS-5) Interoperability Testbed* (OGC 2008) that SOAP and WSDL interfaces have been developed for four foundation OGC interface standards, WMS, WFS-T, WCS-T, and WPS, allowing these services to be integrated into industry standard service chaining tools.

In 2007, the OGC released a Request For Quotation and Call for Participation (RFQ/CFP) to

solicit proposals for the *Canadian Geospatial Data Infrastructure (CGDI) Interoperability Pilot* project (http://www.opengeospatial.org/standards/requests/38). One of the principles of the CGDI is the widespread dissemination of data, which is at the same time managed at or near its source. Data users require authoritative geospatial information, accessible directly from as close as possible to its source, in order to make timely and effective decisions. Evolving the Canadian GeoBase portal to operate in a more distributed fashion and making maintenance transactions more efficient, will help to meet those user requirements.

The CGDI project focuses on three vector-based data themes: geographic name, national road network and administrative boundaries. The functional scope of this project includes investigations in the following areas: access by users to closest-to-source data, transactional updates exchanged between data suppliers and GeoBase, and the use of distributed services architecture to support end-user online applications. The project includes participants from provincial and federal agencies and from the private sector. Agencies participate at different levels, some work with private sector partners while others will use their existing infrastructures to provide access to manage and disseminate data. While this project, strictly speaking, does not involve Grid computing, according to the RFQ/CFP it does aim to take Web services one step further by introducing transactions into the Web Feature Server. Like Compartimos, this project works with vector data in an SDI environment. A live demonstration of the project, available online at http://www.ogcnetwork.net/cgdi, showed the benefits of access to a CGDI distributed network of federal, provincial and territorial data servers (14 in total) and highlighted:

1. access to place names, roads and municipal boundary data from a distributed network of federal, provincial & territorial servers (14 in total);
2. direct updates of data in provincial servers; and
3. a demonstration of an emergency response scenario

In 2007 the OGF and OGC signed a memorandum of understanding (MoU) to collaborate (GridToday 2007). The OGC Web Processing Service (WPS) was chosen as a starting point with the following initial goals:

- Integrate the WPS with a range of “back-end” processing environments to enable large scale processing as an application driver for both grid and data interoperability issues.
- Integration of the WPS with workflow management tools.
- Integration of OGC catalogues and data repositories with grid data movement tools such as GridFTP.

The goal is to enhance operational hurricane forecasting, location-based services and anything to do with data on a map, which naturally includes address data. The two organizations hope that from the mutual understanding of technical requirements and approaches, other opportunities and capabilities will emerge. The announcement of the MoU was published in November 2007 and thus the collaboration is still in its very initial stages. However, the collaboration proves that on an international level, the geospatial community is increasingly interested in utilizing grid technology as solution to its problems, while the grid community has found another community that can benefit from its technology. The results from the work in this dissertation will provide valuable input towards the mutual understanding of requirements and approaches in the Grid and geospatial communities, which the OGF/OGC collaboration seeks to build.

Rajabifard *et al.* (2005) acknowledge that new developments in database management software, including Grid computing technologies will change the way in which data is stored and maintained. However information about actual application of grid technology in SDIs is limited. One such project that incorporates Grid computing technology into an SDI is the recently launched *Geodateninfrastruktur-Grid* (GDI-Grid) project (http://www.d-grid.de/index.php?id=398&L=1), which is part of D-Grid, a long-term German strategic initiative in Grid computing. *Geodateninfrastruktur* is the German word for spatial data infrastructure and the project aims to find solutions for the efficient integration and processing of geographic data within geographic information systems (GIS) and spatial data infrastructures. As such it is the aim of GDI-Grid to develop standardized SDI services that can be used by a wide range of users and/or applications and thereby opening up access to existing SDI resources. The project was launched in July 2007 and to date there no publications have been listed on the project's website. From an OGF-22 workshop report (OGF 2008b) it is evident that members of the GDI-Grid are involved in the OGC-OGF Collaboration. GDI-Grid operates in an SDI environment and from that point of view is similar to Compartimos, but from the information available on the project, the GDI-Grid does not seem to include address data specifically.

52°North Initiative for Geospatial Open Source Software GmbH (http://52north.org/) is an international research and development company with the mission to promote the conception, development and application of free open source geo-software for research, education, training and practical use. One of their focus areas is geo-processing, which allows for the standardized and web-based processing of geographic data. They have developed a number of web services based on the OGC's Web Processing Service (WPS) interface standard. First reports on the grid-enablement of WPS, for example, Baranski (2008), have been published on their website. *52°North* are also involved in the OGC-OGF Collaboration according to the OGF-22 workshop report (OGF 2008b).

From the work that is discussed in this section, it is clear that the research community, as well as industry, recognizes the importance of grid computing for SDIs and geospatial data in general. The 52°North Initiative and the GDI-Grid project, as well as the MoU between the OGF and the OGC, are proof of this recognition, but the projects are still in their initial stages and results are not yet available. While ESG and GEON are examples of existing spatial data grid implementations, their purpose is slightly different to that of a data grid in an SDI environment that incorporates multiple local authorities.

In summary, the related work that was discussed here confirms that the approach in this dissertation of the data grid as enabler for SDI data sharing is innovative and new, and it proves that the work is extremely relevant at this point in time in both the Computer Science and the GISc disciplines. The work is also unique because Compartimos is designed for address data.

Chapter 4 Compartimos, a reference model for an address data grid in an SDI

4.1 Introduction

A *model* is a simplified representation of a system or phenomenon that is used in the sciences to describe a system, often mathematically (Cambridge University Press 2007, Oxford University Press 2007a, Dictionary.com 2008); a *reference model* is an abstract framework for understanding significant relationships among the entities of some environment (OASIS 2008). In this chapter the *Compartimos reference model*, developed by the author, is presented. 'Compartimos' is the Spanish word for 'we share' and the Compartimos reference model gives an abstract representation of the essential components and their relationships that are required to *share* address data on a data grid in an SDI environment. Compartimos serves to analyze the problem space of data grids and SDIs by addressing a very specific problem in these areas (sharing address data on a data grid in an SDI) and provides valuable feedback about the usability of the general models (data grids and sharing data in an SDI) in this specific area of interest. Compartimos also has a clarifying purpose, because it is one of the first attempts in the joint SDI and data grid problem space, and thus enhances the mutual understanding of these two domains. The two problem spaces respectively represent the two disciplines in this dissertation: the data grid problem space in the Computer Science discipline and the SDI problem space in the Geographic Information Science discipline. This chapter thus touches on both disciplines.

The remainder of this introductory section of this chapter is structured as follows: section 4.1.1 clarifies the purpose of Compartimos as a reference model and describes how it contributes to research on data grids and SDIs, and section 4.1.2 describes how Compartimos is presented in the remaining sections 4.2 to 4.6 of this chapter.

4.1.1 The purpose of Compartimos

In the OASIS Service Oriented Architecture Reference Model, a reference model is defined as an *abstract* framework for understanding significant *relationships* among the *entities* of some *environment*, and for the development of consistent standards or specifications supporting that environment (OASIS 2008).

- A reference model is *abstract*, i.e. it does not describe actual things but rather it describes representations of things, or concepts.
- A reference model includes both *entities* (abstract things) *and relationships* (interaction between the things); entities on their own are not sufficient.
- A reference model applies to a specific *environment* or problem space (it is not an attempt to describe or understand everything), which needs to be clearly defined.

Other examples of reference models are the CIDOC Conceptual Reference Model (CRM) for cultural heritage documentation (ISO 21127:2006), the Open Systems Interconnection (OSI) Reference Model that describes computer network architecture (ISO/IEC 7498:1994), Reference Model for Open Distributed Processing (RM-ODP), a reference model for distributed processing (ISO/IEC 10746:1998) and the Spatial Reference Model (SRM) *(*ISO/IEC 18026:2006*)* for applications whose spatial information requirements overlap the scope of the work of more than one ISO technical committee.

Olivier (1999) describes the purpose of a model in research as follows: during the early stages of research in a particular problem space, a model serves to confirm the existence of the problem and to clarify the problem space. Once a few of these models have been developed, the purpose of a model becomes analytical by addressing a more specific problem in the problem space. From a collection of these models for specific problems, trends are observed and one can derive a general model that caters for most (if not all) the assumptions.

Compartimos relates to three existing reference models:

- OGSA and the OGSA data architecture (OGF 2006, OGF 2007a);
- the ISO/TC211 reference model (ISO 19101:2002*)*; and
- the OGC reference model (OGC 2003).

The purpose of the OGC and ISO/TC 211 reference models is to guide standardization efforts in their respective communities, while OGSA and the OGSA Data Architecture are abstractions of distributed systems and their capabilities for a wide range of applications. Compartimos is also an abstraction of a distributed system and its capabilities, albeit for a very specific problem space: sharing address data in an SDI. The OGC and ISO/TC211 reference models are of interest because they also fall within the geospatial domain. In the following few paragraphs Compartimos is discussed in relation to each one of these three reference models.

The *Open Grid Services Architecture (OGSA)* is a vision of a broadly applicable and adopted framework for distributed system integration, virtualization, and management, and defines a core set

of interfaces, behaviors, resource models, and bindings. OGSA provides an abstract definition and is generic, i.e. not specific in terms of the underlying infrastructure (hardware, operating systems, network protocols, etc.). One of the purposes of OGSA was to 'frame the "Grid" discussion' (OGF 2006), in other words, to define and clarify the problem space. The *OGSA Data Architecture* addresses a specific OGSA capability, namely data management, and provides a high-level description of the interfaces, behaviors, and bindings for manipulating data within the broader OGSA architecture. In terms of data, the OGSA Data Architecture is generic: the term 'data' refers to any data, including a sequence of bytes, files, sets of files, or even structured data such as that found in a DBMS (OGF 2007a). Compartimos is a specialization of the OGSA and OGSA data architecture for a very specific environment (SDIs) and serves to analyze the problem space of data grids by addressing a very specific use case for data grids.

The *ISO/TC 211 reference model* defines the framework for standardization in the field of geographic information (the ISO 19100 series of standards) and sets forth the basic principles by which this standardization takes place. Standardization in ISO/TC 211 is mainly focused on the information and computational viewpoints of the RM-ODP (refer to section 4.1.2 for a description of these viewpoints). An important goal of the ISO 19100 series of standards is to create a framework in which spatial data interchange and service interoperability can be realized across multiple implementation environments. Compartimos provides one such example of spatial data interchange and service interoperability in the specific domain of address data in an SDI environment, while the scope of ISO/TC 211 includes *any* digital geographic information, i.e. it is much wider. By addressing a very specific type of spatial data, Compartimos serves to describe and analyze the domain of digital geographic information. On the other hand, some ISO 19100 standards are potential technology choices for Compartimos, as described in Chapter 5.

The *OGC reference model* provides a framework for the ongoing work of the Open Geospatial Consortium. The reference model is presented in terms of the five viewpoints of the RM-ODP: the enterprise, information, computational, engineering and technology viewpoints (refer to section 4.1.2 for a description of these viewpoints). The OGC reference model describes the requirements baseline for geospatial interoperability in terms of these viewpoints, and any specification, experiment, other document or work produced by the OGC is related to one of these viewpoints. Once again, Compartimos is a special use case of geospatial interoperability as generally described in the OGC reference model, and thus Compartimos has an analytical function. Some OGC specifications are also discussed as Compartimos technology choices in Chapter 5.

As described earlier (refer to section 2.5), ISO is an international organization and its members are mainly from the public sector, while OGC's members are mostly from the private sector. In general, ISO has broader goals and is working at a level of abstraction above OGC so that the two

efforts complement each other: ISO publishes standards of which quite a few are abstract standards, while OGC publishes implementation specifications.

Compartimos describes how the general models, i.e. the models from OGF, OGC and ISO/TC 211, can be applied to the specific problem area of address data sharing in an SDI, thereby providing valuable feedback about the usability of the general models in a specific problem area. But Compartimos also has a clarifying purpose, because it is one of the first attempts in the joint SDI and data grid problem space, and thus enhances the mutual understanding of these two domains. To understand this clarifying purpose, one could compare Compartimos to the Open Systems Interconnection Basic Reference Model (OSI Reference Model, which is an abstract description for layered communications and computer network protocol design that has never been implemented and has been superseded by newer IEEE and IETF protocol developments. However, it is still considered as good introductory study material for computer networks and therefore included in textbooks (Tanenbaum and van Steen 2007, Colouris *et al.* 2005). While implementations of Compartimos are indeed possible (refer to the proof of concept implementation described in Chapter 5), the clarifying role of Compartimos is manifested in the reference model itself and does not require an implementation.

4.1.2 Presentation of Compartimos

In this chapter Compartimos, a reference model for an address data grid in an SDI, is presented by means of the five viewpoints prescribed in the RM-ODP (ISO/IEC 10746:1998). The RM-ODP family of recommendations and international standards defines essential concepts necessary to specify open distributed processing systems from five prescribed viewpoints and provides a well-developed framework for the structuring of specifications for large-scale, distributed systems. The RM-ODP is a joint effort by ISO/IEC (International Organization for Standardization/International Electro-technical Commission) and ITU-T (International Telecommunication Union's Telecommunication Standardization). The rapid growth of distributed processing has led to the widespread adoption of the RM-ODP, and the ISO/TC 211 and OGC reference models, for example, are described in terms of RM-ODP viewpoints.

An RM-ODP viewpoint is an abstraction that yields a specification of the whole system related to a particular set of concerns. The five viewpoints defined by RM-ODP have been chosen to be both simple and complete, covering all the domains of architectural design. These five viewpoints are:

- the *enterprise viewpoint*, which is concerned with the purpose, scope and policies governing the activities of the specified system within the organization of which it is a part;
- the *information viewpoint*, which is concerned with the kinds of information handled by the system and constraints on the use and interpretation of that information;

- the *computational viewpoint*, which is concerned with the functional decomposition of the system into a set of objects that interact at interfaces - enabling system distribution;
- the *engineering viewpoint*, which is concerned with the infrastructure required to support system distribution; and
- the *technology viewpoint*, which is concerned with the choice of technology to support system distribution.

Compartimos is presented in terms of the RM-ODP viewpoints because the viewpoints provide a simple yet complete overview of a distributed system and because they have been used successfully to describe other reference models. The viewpoints relate to Compartimos as follows:

- The *enterprise viewpoint* (section 4.2) is concerned with the purpose, scope and policies governing the activities of the address data grid in an SDI. The characteristics of the SDI environment in which address data is produced, maintained and used are described and set the stage for the remaining viewpoints on Compartimos.
- The *information viewpoint* (section 0) is concerned with the kinds of information handled by the address data grid and the constraints on the use and interpretation of that information. In Compartimos there are two types of information: the address data itself and the metadata required for the operation of the address data grid.
- The *computational viewpoint* (section 4.4) is concerned with the functional decomposition of the address data grid into a set of objects that interact at interfaces, thereby enabling a single virtual address dataset. In line with current trends in grid computing, Compartimos has a service-oriented architecture. The purpose and capabilities of the essential services required to implement an address data grid are described, as well as the way in which these services interact with each other.
- The *engineering viewpoint* (section 4.5) is concerned with the infrastructure required to support the interaction of the objects that enable the virtual address dataset and therefore includes details about the deployment options for the reference model objects.
- The *technology viewpoint* is concerned with the choice of specific technologies in support of implementation of the address data grid and is discussed in Chapter 5, together with the implementation of Compartimos.

Section 4.6 of this chapter concludes with a discussion of the four viewpoints on Compartimos that are presented in this chapter.

4.2 Enterprise viewpoint

In this section the *enterprise viewpoint of* Compartimos is presented. It is concerned with the purpose, scope and policies governing the activities of an address data grid in an SDI. The SDI environment has some unique characteristics that influence the way in which a data grid can be implemented and therefore it is necessary to take these characteristics into consideration as part of the enterprise viewpoint. Section 4.2.1 addresses the scope and purpose of the address data grid, and section 4.2.2 describes some high-level use cases to illustrate this purpose. A virtual organization (VO) comprises the set of individuals and/or institutions sharing data in a data grid and in section 4.2.3 a VO for an address data grid in an SDI is described. The SDI environment in which address data is produced and maintained has already been described in Chapter 2 and is summarized again in section 4.2.4. The enterprise viewpoint described in this section sets the stage for the remaining viewpoints on Compartimos.

4.2.1 Scope and purpose

An address data grid in an SDI has to provide a *non-trivial service* of coordinated access to distributed heterogeneous address data *resources* that are not subject to centralized control. These are data resources at various local authorities as they typically occur in an SDI environment. The data grid makes use of *standard, open protocols and interfaces*, unless they are not (yet) available.

The purpose of the address data grid is threefold. Firstly, the goal is to make a number of individual address datasets, each under the control of a different institution, available as a *single virtual address dataset* that spans a larger area. This virtual address dataset is created dynamically from the different geographically distributed heterogeneous data resources. For example, in a national SDI, the address data grid would provide access to a virtual national address dataset that spans the whole country and comprises the local address datasets of individual local authorities. In an international SDI such as INSPIRE for Europe, the address data grid would provide access to a virtual international address dataset that spans the whole of Europe and comprises the address datasets of individual countries. Access to the virtual address dataset should be provided in a uniform way, even though individual address data providers produce and maintain address data in their own proprietary vendor-specific format according to their own specific data model, semantics and business logic. Access to the data grid should not be restricted to specific platforms of operating systems, programming languages, geographic information system vendors, or data formats.

Secondly, the aim is to make this virtual address dataset available *to as wide an audience as possible* which implies that access services to the virtual address dataset have to be based on standardized and open interfaces and protocols, and that the data grid has to be scalable so that the number of users can continue to grow. Access to the database should include both bulk up- and

downloads, as well as high volumes of individual address queries.

Thirdly, third party organizations should be able to *provide services on top of the single virtual address dataset*. Address data only really becomes useful when it is integrated into other services such as routing, address capturing, geocoding, and mapping. Therefore, apart from creating an infrastructure that gives access to the 'raw' address data, an SDI should also provide the infrastructure to enable third party services for routing, address verification, geocoding, mapping, etc. on top of the single virtual address dataset.

The data grid described in this dissertation is limited to address data but serves as an example for other spatial datasets such as points of interest, traffic lights, man holes, cadastral information and road networks that are also produced and maintained by individual local authorities. Compartimos is a profile (or customization) of the OGSA data architecture for address data in an SDI. In Table 5 an address data grid in an SDI environment of Compartimos is compared to the table of data grid implementations that was presented in Chapter 3 .

Table 5. Compartimos compared to the data grid implementations described earlier in Chapter 3

	Compartimos	LIGO	Earth System Grid	e-DiaMoND	GEON
Application domain	Spatial data infrastructures (SDIs)	Physics and astronomy	Climate modeling	Breast cancer treatment	Earth sciences
Region	Regional, national or international	United States	United States	United Kingdom	Northern America
Number of data sites	Ranging between ten (small region) and a few thousand (local authorities in a country)	Two	Around 10 centers and laboratories	Scalable to 90+ Breast Care Units (BCUs) in the UK	3 GEON data nodes, 15 GEON points of presence
Total data volume	Between 2.5GB and 5TB, depending on the size of the region and the size of the individual address records (conservative estimate)	One terabyte per day, ca. 365 terabytes per year	250 terabytes until 2006, ca. 70 terabytes per year	Estimated 480 terabytes per year, when fully operational	Each data node can store 4 terabytes of data
Metadata	Based on *ISO 19115 – Geographic information - Metadata*, stored in a relational database	Descriptive metadata about the data in the files (in a relational database)	Climate model metadata (in a relational database)	Patient data and metadata on image files (in a relational database)	Metadata about data made available by providers (in a relational database)
Format of data resources	Proprietary GIS file such as .SHP files, relational databases such as Oracle and SQLServer storing spatial data.	Files with data from the LIGO detector	Files containing climate research data	Image files with mammography	Relational data, ESRI .SHP files, LiDAR
Size of individual data item	The size of an individual address record ranges between 5K and 10K (conservative estimate)	1-100 megabytes per file	Unknown	Estimated 75 megabytes per image file	Varies considerably, depending on what a user uploads
Number of data items	Ranging between 500,000 (small region) and approximately 500,000,000 (international region), depending on the size and address density of the region	More than 40 million files	Millions of files	1000 cases	Around 4,500 (searching the portal on March 2008)
Interaction	Portal, as well as web services	Portal www.ligo.org	Portal www.earthsystemgrid.org	Service registry and Web services	Portal www.geongrid.org, as well as Web service registry

4.2.2 High-level use cases

To illustrate the threefold purpose described in the previous section, this section presents three high-level use cases for Compartimos:

1. a *simple data request* (purpose: single virtual address dataset);
2. an *iterative data request* (purpose: make the data available to as wide an audience as possible); and
3. a *third party service request* (purpose: provide services on top of the single virtual address dataset).

Figures 17-19 below illustrate the three use cases and a brief description of each use case is given. The use cases are refined and discussed further in the remainder of the chapter as part of the other viewpoints of Compartimos.

Simple data request. The user specifies a filter such as a bounding box, and all the address data that is available in the data grid satisfying the filter is returned. For example, a simple data request is executed when a mapping application requests addresses that are to be displayed within the current zoom scale of its map. This use case illustrates Compartimos' goal to make individual address datasets appear as a single virtual address dataset, but it also contributes to making the address data available to as wide an audience as possible. In the SDI context, this use case represents the search for address data resources at local authorities, returning from these data resources any address data that matches the input filter.

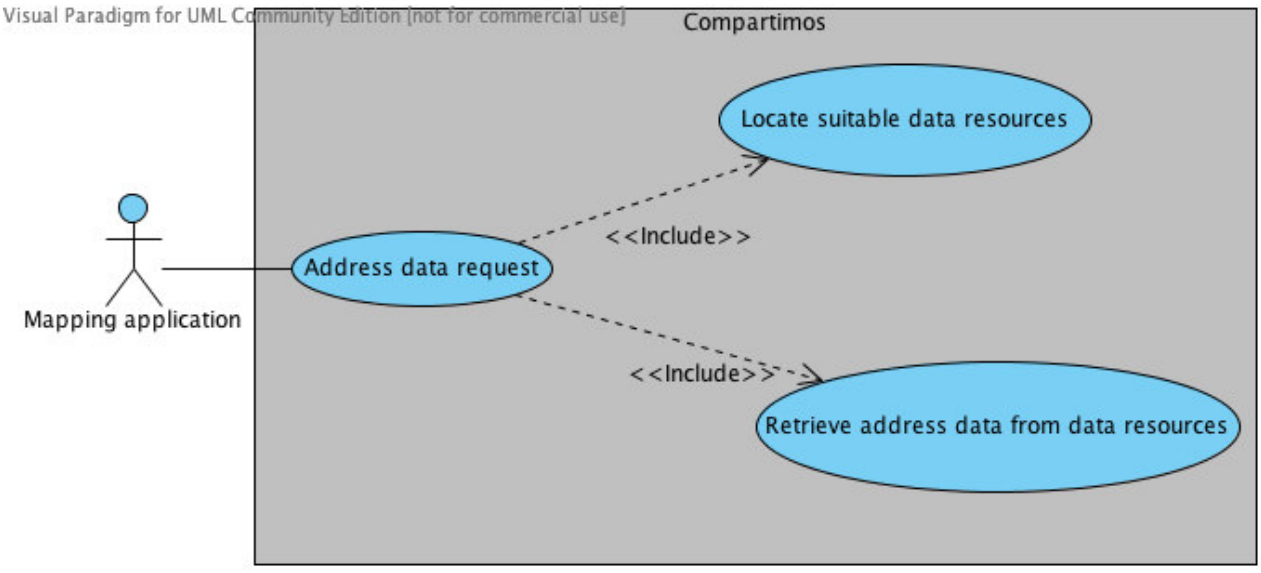


Figure 17. Simple data request (use case)

Iterative data request. In the iterative data request the level of detail of the requested address data is increased iteratively with each subsequent request. This type of request makes address data available to as wide an audience as possible, and is used to allow a user to select a valid address from dropdowns, for example: when capturing the residential address the dropdowns guide a user in selecting an address that is valid by first presenting a list of addressing systems, such as a street address, a site address etc., and then a list of, for example, provinces, municipalities, suburbs, streets and so on.

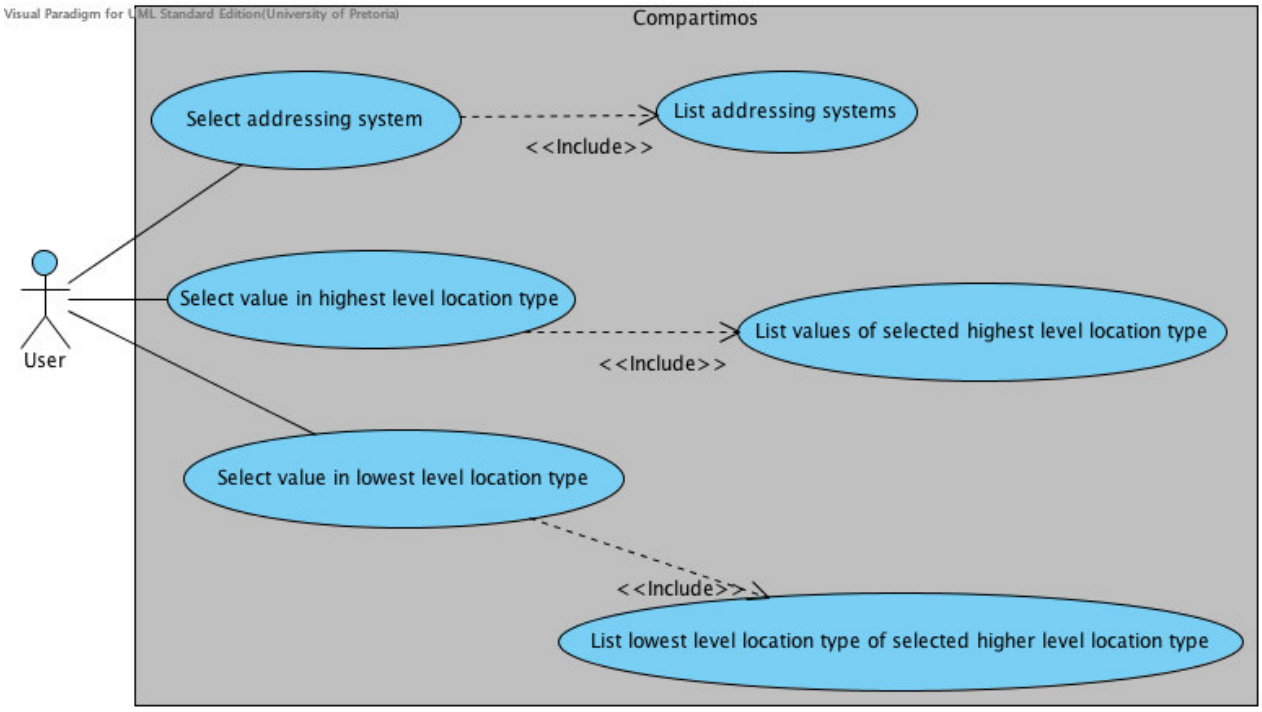


Figure 18. Iterative data request (use case)

To better explain this use case, the following steps describe how a user would interact with an application, giving actual values as examples:

1. The user is presented with a dropdown list of addressing systems, such as 'SANS 1883 street address type', 'SANS 1883 intersection address type', etc.
2. The user selects the 'SANS 1883 street address type' in the dropdown.
3. Next, the user is presented with a list of values from the highest-level location type of the 'Street address type', i.e. the Province. In other words, the user is presented with the list of provinces of from South Africa, i.e. Eastern Cape, Free State, Gauteng, etc.
4. The user selects a province, e.g. Gauteng.
5. Next the user is presented with a list of municipalities (the next highest level location

type) in Gauteng, i.e. City of Tshwane Metropolitan Municipality, Emfuleni Local Municipality, Lesedi Local Municipality, etc.

6. The user selects a municipality and the above process continues until he user is presented with values from the lowest level location type, the Street Number.

Service request. The user requests a route between two or more addresses. Only the first step is really part of the data grid, when the addresses have to be converted into coordinates. During the subsequent two steps the coordinates are first 'snapped' to the closest nodes in a street network (note this data is external to the address data grid) and then a route between the coordinates is calculated. This use case illustrates how third party services can be provided on top of the single virtual address dataset, representing the third purpose described in the previous section.

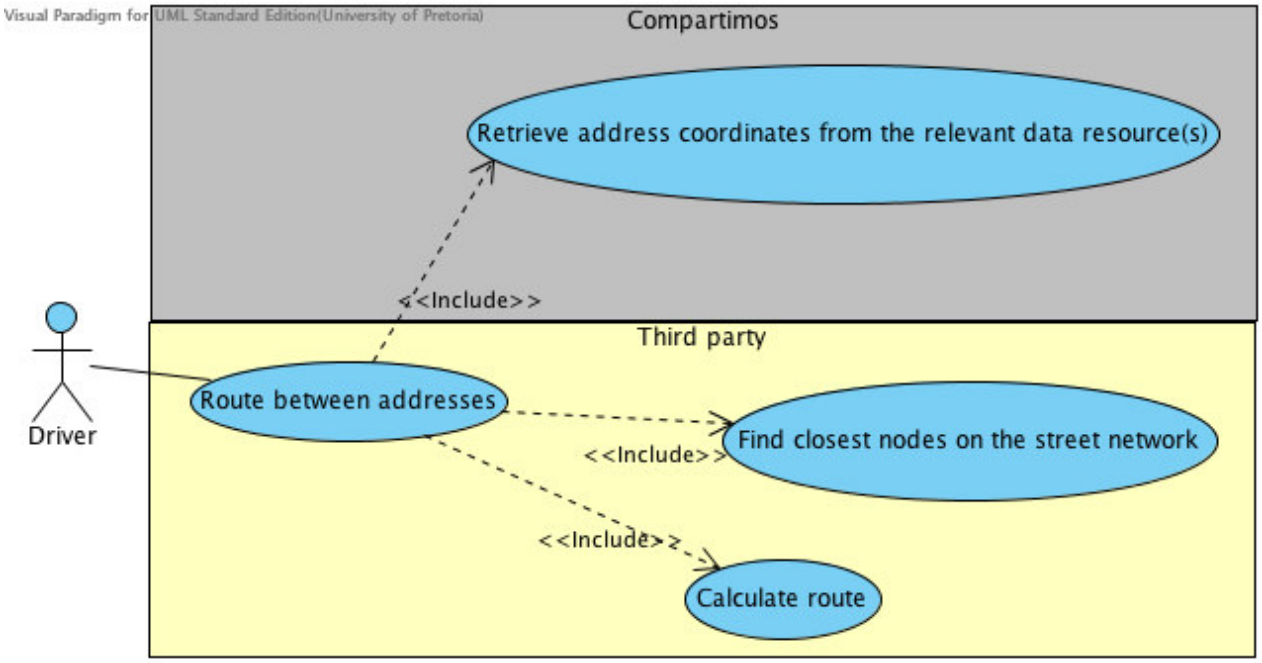


Figure 19. Service request (use case)

4.2.3 Virtual organization (VO)

The virtual organization (VO) is an important concept in a data grid and therefore in this section the VO characteristics of an address data grid in an SDI are described. In general, a VO comprises a set of individuals and/or institutions having direct access to computers, software, data, and other resources for collaborative problem solving or other purposes. VOs are a concept that supplies a context for operation of a Grid that can be used to associate users, their requests, and a set of resources. The sharing of resources in a VO is necessarily highly controlled, with resource providers and consumers defining clearly and carefully just what is shared, who is allowed to share, and the

conditions under which sharing takes place (OGF 2007c). In the specific case of this dissertation, a VO comprises a set of individuals and/or institutions having direct access to address data from various address data sources (the resources) that is presented to the users as a seamless, single virtual address dataset (the purpose), and the individuals and/or institutions participate in the following VO member roles or capacities:

1. The *address data provider* is the institution that publishes the address dataset on the data grid. This could be the custodian or owner of the data, such as a local authority, but can also be an appointed distributor of the data, such as a consultant acting on behalf of a local authority. The address data provider produces new releases of the data and defines what data is shared, who is allowed to share the data, and the conditions under which the sharing takes place (in agreement with the owners of the data, of course).
2. The *address data host* is the institution that provides the required resources to host the dataset on the data grid. For this, it has to provide an implementation of the uniform interface to the underlying address dataset, as well as a hosting environment for the interface and the data itself. Thus, the host makes the data available on the data grid in a uniform way. The data host could be the same institution as the data provider, or it could be a third party, such as an ISP, which provides the hosting service to the data provider and uses an implementation of the uniform interface that is provided by any arbitrary institution (including an open source project team), as long as it conforms to the uniform interface.
3. The *node host* is the institution that provides the resources to host a point of presence in the data grid. In data grid literature the node is sometimes referred to as a point of presence (e.g. in the GEON grid). The node comprises the catalogue and virtual address data services together with optional services for replication, transfer, etc. There are different levels of nodes depending on whether the node hosts the optional services and/or provides additional storage space for uploading address data to the grid.
4. The *address data consumer* is any user (an individual user, an institution or an application) that requests data, whether for mapping, address capturing, routing or otherwise, from the address data grid. In the use cases of Figures 17 and 18 the consumer is represented by the mapping application and customer respectively.
5. The *address-related service provider* is any third party providing address-related services, such as routing, on top of the single virtual address dataset. By definition, the service provider is also an address data consumer.
6. The *address-related service consumer* is any user (an individual user, an institution or an application) that consumes an address-related service such as a routing service provided on

top of the data grid. In the use case of Figure 19 the driver is the service consumer. A VO member could be both a data consumer as well as a service consumer.

An institution can adopt more than one role in the VO. For example, a small local authority might opt to not be a data provider at all, but to appoint a consultant as its data provider, which in turn outsources the data hosting to a third party; a medium-sized local authority might be a data provider, data host and node host; and a larger local authority might be a data provider, data and node host, as well as an address-related service provider. Also important to note is the contribution of VO members that (merely) host one or more nodes allowing the grid to be scaled up, even more so if these nodes also allow address data to be replicated to their sites. Table 6 provides examples of some of the combinations of roles.

In its simplest form the VO has members that are data providers, data hosts, node hosts and external data consumers. Figure 20 shows how these members could be distributed among different organizations, each representing a different administrative domain. A VO in the address data grid can be short-lived, for example, for the duration of a specific disaster relief operation; or long-term, for example, for the verification of residential addresses of new customers when applying for a financial account.

Table 6. Member roles in a VO

Data provider	Data host	Node host	Service provider	Example institution
✓				Small local authority that only produces and maintains the address data
	✓			Consultant that hosts the data on behalf of a small local authority
	✓	✓		Private company providing a hosting service to a small local authority
✓	✓			Small local authority that provides and hosts its own data
✓	✓	✓		Medium-sized local authority that provides data and hosts the data and a node
✓	✓	✓	✓	Metropolitan local authority that provides data, hosts the data and a node and also provides address-related services
			✓	Private company that provides address-related services on top of the address data grid
		✓		National authority that hosts one or more nodes and thereby increasing the scalability of the address data grid
	✓	✓	✓	National authority that provides data and node hosting services to smaller authorities, as well as address-related services
✓	✓	✓	✓	National authority that provides data (e.g. the post office), hosts data and a node, as well as address-related services

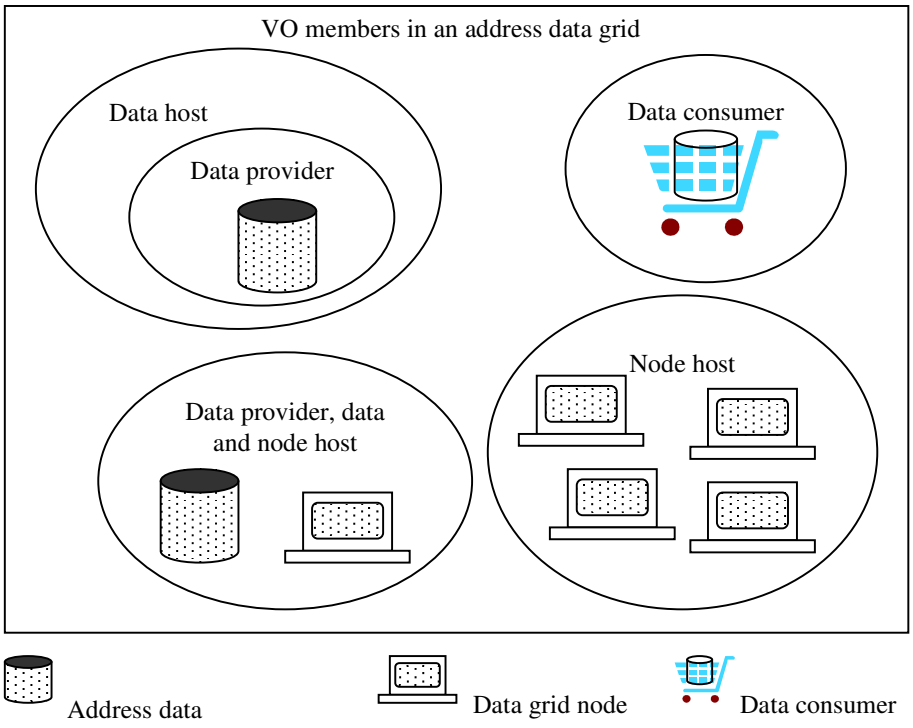


Figure 20. The VO distributed across different administrative domains (represented by the ovals)

4.2.4 SDI environment of address data

Compartimos is intended for address data in an SDI environment, and in this section this environment and the policies and constraints that go along with it are described. There are three major role players in a national SDI environment: national or regional government, local government, and the private sector. Ideally, national government should play a strategic role, while local governments are responsible for the production and maintenance of address data in line with the national strategy. Both local government, as well as national government, are users of address data, but the role of the private sector is increasing as companies are starting to incorporate address data into, for example, corporate databases where a customer is linked to an address. Apart from local governments, there are other producers of address data, as illustrated in section 2.2.2 earlier. In an international SDI environment, which Compartimos aims to accommodate, the number of role players increases along with the level of heterogeneity of address data producers and users. As a result, Compartimos has to cater for the heterogeneous environments in which address data is produced, maintained and consumed, characterized by the following:

- *Heterogeneous platforms.* Address data is produced, maintained and consumed on different operating systems, DBMSs and programming language platforms.

- *Heterogeneous data models.* Address data is modeled according to each data provider's own specific needs, resulting in syntactic and semantic differences between the data models.
- *Multiple address data producers.* There could be more than one producer of address data for a specific area. These could include both producers that have been officially assigned as custodians for an area (such as local governments), as well as unofficial producers (such as utility or private companies) who assign address data for their own purpose and use, or for resale of data products.
- *Varying coverage areas.* Producers work with coverage areas of varying sizes, depending on their area of interest. These coverage areas range from the whole country to a province, local authority or even a single suburb.
- *Multiple decentralized sources of address data.* There are many decentralized sources of address data, and these sources are continuously updated.
- *Data access management.* Owners of address data need to be able to specify who can access which parts of their data in which manner, and require knowing when, by whom and how their data was accessed.

Figure 21. Potential for address ambiguity

The perception of the person on the street – the user – of an address is often very different from the officially assigned address. The inconsistent use of place names is a good example of this perception 'problem'. Place names are the cause of ambiguity in an address when the colloquial use of a place name differs from the officially assigned place name. Since place name boundaries are not physically fenced off, these boundaries are easily misinterpreted (Coetzee and Cooper 2007b). Figure 21 illustrates such a potential misinterpretation since the boundary between the suburbs of 'Murrayfield' and 'Die Wilgers' runs along the centre of Rubida Street for only a part of the street. Thus an address in Rubida Street could be in one of three suburbs: Lynnwood Ridge, Murrayfield or Die Wilgers, but the person on the street it is not obvious to which suburb a particular address belongs. Multiple producers of address data are another source of ambiguity in place name usage. The address data grid has to provide a means of resolving these ambiguities.

4.3 Information viewpoint

The *information viewpoint* is concerned with the kinds of information handled by Compartimos along with the constraints on the use and interpretation of that information. The information viewpoint of Compartimos is presented in two sections: section 4.3.1 deals with the representation of address data itself; and section 4.3.2 deals with the catalogue of metadata containing information about addressing systems, address datasets, data providers, data and node hosts, and service providers, along with the address-related services that they offer. In the OGSA Data Architecture Scenarios document, three potentially complex steps of the data integration scenario (OGF 2007b) are described and these apply to Compartimos as well:

1. *Data discovery*: if the locations of the address datasets are not already known, they have to be discovered via registries or directories of address data sources.

2. *Schema mapping*: the address data must be understood and presented in a uniform manner, requiring the capability to map between the different schemas describing the address data.

3. *Data consolidation*: differences in the format or structures of the address data from disparate heterogeneous environments may require transformations to a single address data format so that the disparate data can be comparable.

This chapter deals with all three of these steps to ensure that data integration in Compartimos can be achieved: section 4.3.1 on address data describes how *schema mapping* is done; section 4.3.2 on the address data catalogue deals with the information that is required for *data discovery*; and *data consolidation* is addressed under the computational viewpoint in section 4.4, mainly as part of the VirtualAddressDataService in section 4.4.6.

4.3.1 Address data

In order to create a single virtual dataset of address data, the address data from multiple producers must be understood and presented in a uniform manner. This requires both syntactic as well as semantic harmonization of the heterogeneous sources of address data, ideally accomplished by a standardized data model for address data. Standards for national address data exist or are under development in a number of countries, such as SANS 1883 (draft), *Geographic information – South African address standard* in South Africa, *AS*/NZS 4819, *Geographic information – rural and urban addressing* in Australia and New Zealand and BS7666, *Spatial datasets for geographic referencing* in Britain. These national standards could be employed to establish an address data grid on a national level. For an international address data grid, however, an overarching international standard that allows address data exchange across national borders is required. Two such standards exist: UPU-S42 by the Universal Postal Union (UPU S-42 2006)), and the address standard produced by the Customer Information Quality (CIQ) committee of OASIS (OASIS CIQ 2007). However, the UPU standard is limited to postal addresses only, while the OASIS standard has some shortcomings in terms of geographic data as described by Coetzee *et al.* (2008b).

Thus, it was necessary to develop a novel address data model for Compartimos that overcomes the shortcomings in the UPU and OASIS address standards, but enabling international address data exchange. This novel data model is based on the three principles listed below, and discussed in more detail in the subsequent paragraphs of this section. These three principles are in line with the goal of using existing standards where they exist (refer to the first paragraph of section 4.2.1).

1. There are different types (or classes) of addresses.
2. Each type of address can be described in terms of an addressing system.
3. An addressing system is a specific class of spatial reference system by geographic identifiers, as described in *ISO 19112 - Geographic information - Spatial referencing by geographic identifiers*.

Firstly, addresses can be grouped into different types, depending on how their contents is structured. The notion of address types is found in national standards for address data, such as SANS 1883, AS/NZS 4819 and the draft US street address standard (Wells *et al.* 2008), each describing different types of addresses based on the contents and structure of an address. SANS 1883, for example, defines twelve types of addresses for South Africa: the Street Address, Building Address, Site Address, Intersection Address, Landmark Address, SAPO Box address, SAPO Street Address, SAPO Site Address, SAPO-type Village Address, SAPO Post Restante Address, Farm Address and Informal Address types. To illustrate the concept of address types, Table 7 lists sample addresses from South Africa together with the address type for each address.

Table 7. South African sample addresses

Address	Contents	SANS 1883 Address Type
45 Marais Street, Rustenburg	Street number, street name, and place name	Street Address
Corner Eagles Drive and Dunn Street, Hillcrest	Intersecting street names and place name	Intersection Address
Parliament, Cape Town	Landmark name and place name	Landmark Address
59 Gannabos Street, Val de Grace, 0184	Street number, street name, place name and postcode	SAPO Street Address
Corner of Festival and Schoeman Streets, Hatfield	Intersecting street names and place name	Intersection Address
Voortrekker Monument, Pretoria	Landmark name and place name	Landmark Address
Spaza shop opposite the taxi rank in Tsamaya Road, Mamelodi	Informal reference and place name	Informal Address
PO Box 10965, Garsfontein, 0181	Post box, place name and postcode	SAPO Box Address
77 Chopin Street, Constantia Park	Street number, street name, and place name	Street Address
14 Castle Pine Crescent, Silver Lakes, 0081	Street number, street name, place name and postcode	SAPO Street Address
'My Farm' sign approx. 10km out of town next to the blue gum plantation, Kimberley Road, Bloemfontein	Farm reference, road name and place name	Farm Address
PO Box 11800, Silver Woods, 0080	Post box, place name and postcode	SAPO Box Address

For simplicity reasons, the number of address types in the Compartimos address data model is restricted, but the advantage of using address types is that all types do not have to be known in advance. Additional address types can be added by individual countries at a later stage without having to change the address data model. In other words, the model provides the meta-language to describe address types.

```
StreetAddress = StreetIdentifier, Locality

StreetIdentifier = [CompleteAddressNumber | StreetNumberRange],
                   CompleteStreetName

CompleteStreetName =
  StreetNameAndType, [[StreetNameDirectional], StreetNameModifier]
| SubStreetNameAndType, [[StreetNameModifier], StreetNameDirectional]
| [StreetNameModifier], SubStreetNameAndType, [StreetNameDirectional]
| [StreetNameDirectional], SubStreetNameAndType, [StreetNameModifier]

Locality = PlaceName, [Town], [Municipality], [Province],
[SAPOPostcode],
```

Figure 22. EBNF for the SANS 1883 Street Address type (SANS/CD 1883-1 2008)

Secondly, for each address type there is an addressing system that describes how the elements of an address are combined to form a valid address, i.e. a system according to which addresses are assigned – the definition for an *addressing system* provided in Chapter 2. In SANS 1883, these addressing systems are defined in terms of Extended Backus Naur Form (EBNF), as can be seen in Figure 22 for the Street Address Type. The US address standard makes use of an EBNF-like notation, while the British address standard uses diagrams that are based on the *Structured Systems Analysis and Design Method* in accordance with British standard BS 7738-1 for logical modeling. The Compartimos address data model is illustrated by means of the Unified Modeling Language (UML).

Thirdly, an addressing system is a specialization of a spatial reference system by geographic identifiers, as described in ISO 19112. The British address standard, BS 7666, is also based on this principle, which was deliberately included in the novel address data model of Compartimos to show that the South African address types can also be modelled based on this principle. This is an indication that ISO 19112 could form the basis for an international address standard from the geographic information community, if it is suitably revised to include, amongst others, location type combinations to represent, for example, intersections.

According to ISO 19112 a *spatial reference* is a description of position in the real world (such as an address) and a *spatial reference system* is a system for identifying position in the real world (such as an addressing system). A *geographic identifier* is a spatial reference in the form of a label or code (such as a place or a street name or an address) that identifies a location. A *spatial reference system using geographic identifiers* is a system for describing positions in the real world with labels or codes and comprises a related set of one or more *location types* that may be related to each other through aggregation or disaggregation, possibly forming a hierarchy. A *gazetteer* is a directory of *instances of location types*.

An *address* is a spatial reference in the form of a hierarchical combination of geographic identifiers. Note that there is a one-to-many relationship between an address and a location type (because an address is a combination of location instances, each of a different location type), while there is a one-to-one relationship between a location instance (name of a place) and a location type. An *addressing system* is a spatial reference system using addresses for describing position in the real world. It comprises a related set of one or more *location types* that usually form a hierarchy. Note that there is a composition association between location types and an addressing system (black diamond: strong aggregation because location types such as street numbers do not have a lifetime of their own), and not an aggregation association (white diamond: weak aggregation where one location type ‘belongs to’ or ‘is part of’ another), as defined in ISO 19112 between the location types and a spatial reference system using geographic identifiers. An *address* is an instance of a valid

combination of location types, as allowed by the rules of the addressing system. An *address dataset* is a directory of *addresses*.

Table 8. Concepts and their relationships in ISO 19112 in relation to Compartimos

ISO 19112	Compartimos address data model
Geographic identifier: a spatial reference in the form of a label or code	Address: a spatial reference in the form of a hierarchical combination of geographic identifiers
Spatial reference system using geographic identifiers: a system for describing positions in the real world with labels or codes, comprising a related set of one or more location types that may be related to each other through aggregation or disaggregation, *possibly forming a hierarchy*	Addressing system: a system for describing position in the real world with addresses, comprising a related set of one or more location types that *usually forming a hierarchy*
Gazetteer: a directory of instances of location types	Address dataset (especially for reference purposes): a directory of valid addresses

Table 8 lists concepts from ISO 19112 and their definitions, together with corresponding concepts and definitions in Compartimos This shows that an addressing system in the Compartimos data model is a specific class of spatial reference system using geographic identifiers.

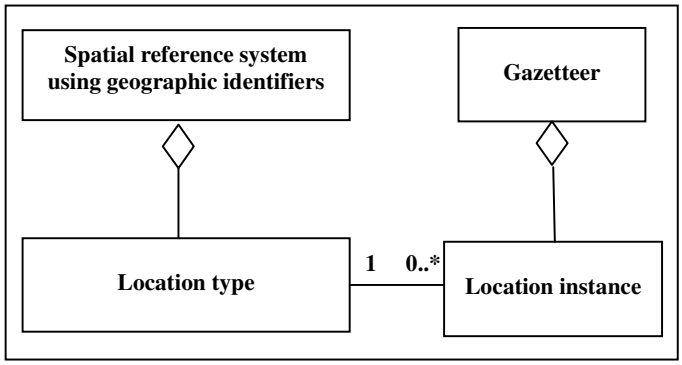


Figure 23. Spatial referencing using geographic identifiers (ISO 19112:2003)

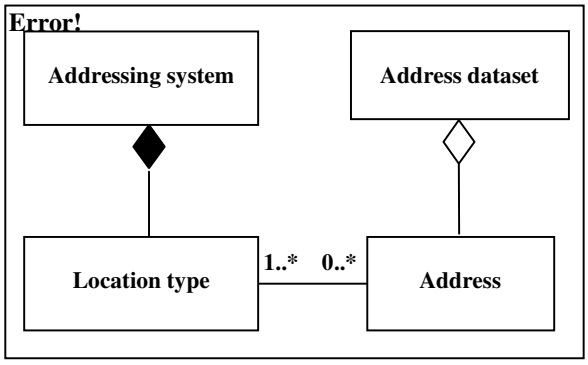


Figure 24. Spatial referencing using addresses (adapted from ISO 19112:2003)

Figure 23 illustrates the relationships among concepts in ISO 19112, while Figure 24 shows relationships among the concepts in the Compartimos address data model. The black diamond in Figure 24 denotes the hierarchical relationship between location types, making them interdependent, while the white diamond in Figure 23 denotes an independently aggregated set of location types forming a spatial reference system using geographic identifiers. Furthermore, the address in Figure 24 links to one or more location type since an address contains multiple geographic identifiers that

are instances of more than one location type, while a location instance in Figure 23 links to one location type only.

An example of spatial referencing using addresses in South Africa, comprises the following hierarchy of location types: *Street Number* > *Street* > *Suburb* > *Municipality* > *Province* > *Country*. This is the addressing system representing the Street Address type in SANS 1883. Figure 25 shows some instances of valid combinations of these location types, thus representing valid street addresses in South Africa.

1083 > *Pretorius Street* > *Hatfield* > *City of Tshwane Metropolitan Municipality* > *Gauteng* > *South Africa*
1083 > *Pretorius Street* > *Hatfield* > *Gauteng*
1083 > *Pretorius Street* > *Hatfield* > *Pretoria*
1083 > *Pretorius Street* > *Hatfield*
Hans Stride Drive > *Faerie Glen* > *Pretoria*
4 > *Church Street* > *Arcadia* > *South Africa*

Figure 25. Valid street addresses according to the SANS 1883 street address type

Figure 54 in Appendix B shows the details of the novel address data model that is based on the three principles explained above. To further illustrate that this data model can be used to describe the SANS 1883 address types, the name and domain of validity attributes defined in ISO 19112 are used in Table 9 to describe the addressing systems of five of the twelve SANS 1883 address types, and also list the location types of the addressing system of the SANS 1883 street address type in Table 10.

Table 9. Descriptions of addressing systems for five of the SANS 1883 address types

Name	Domain of validity	Location types
Street Address	South Africa	Street number, street name, place name, town, municipality, province, country
Site Address	South Africa	Address number, place name, town, municipality, province, country
Intersection Address	South Africa	Street name, intersection street name, place name, town, municipality, province, country
SAPO-type village address	South Africa	House number, village name, SAPO post office name, SAPO street postcode
Informal address	South Africa	Informal reference, place name, town, municipality, province, country

Table 10. Location types of the addressing system of the SANS 1883 street address type

Name	Identifier	Description	Territory of use	Owner	Parent	Child
Street number	Number	Identifies individual dwelling or site	South Africa	Local authority	Street	None
Street name	Name	Thoroughfare providing access to properties	South Africa	Local authority	Place name, town, or municipality	Street number
Place name	Name	Registered or colloquial name for the area/community	South Africa	Local authority or colloquial	Town, municipality, province or country	Street name
Town	Name	More or less coincides with pre-2001municipal boundaries	South Africa	Colloquial	Municipality, province or country	Place name or street name
Municip ality	Name or code	Official municipal boundaries	South Africa	Municipal Demarcation Board	Province or country	Town or place name
Province	Name or code	Official provincial boundaries	South Africa	Municipal Demarcation Board	Country	Municipality, town or place name
Country	Name or code	Country border	South Africa	Municipal Demarcation Board	None	Province, municipality, town or place name

Figure 26 provides a visual presentation of the hierarchical relationships between the location types of the addressing system representing the SANS 1883 street address type. Figure 27 visually presents some examples of valid combinations of instances of these location types, i.e. valid street addresses, such as *1083 Pretorius Street Hatfield Gauteng South Africa*.

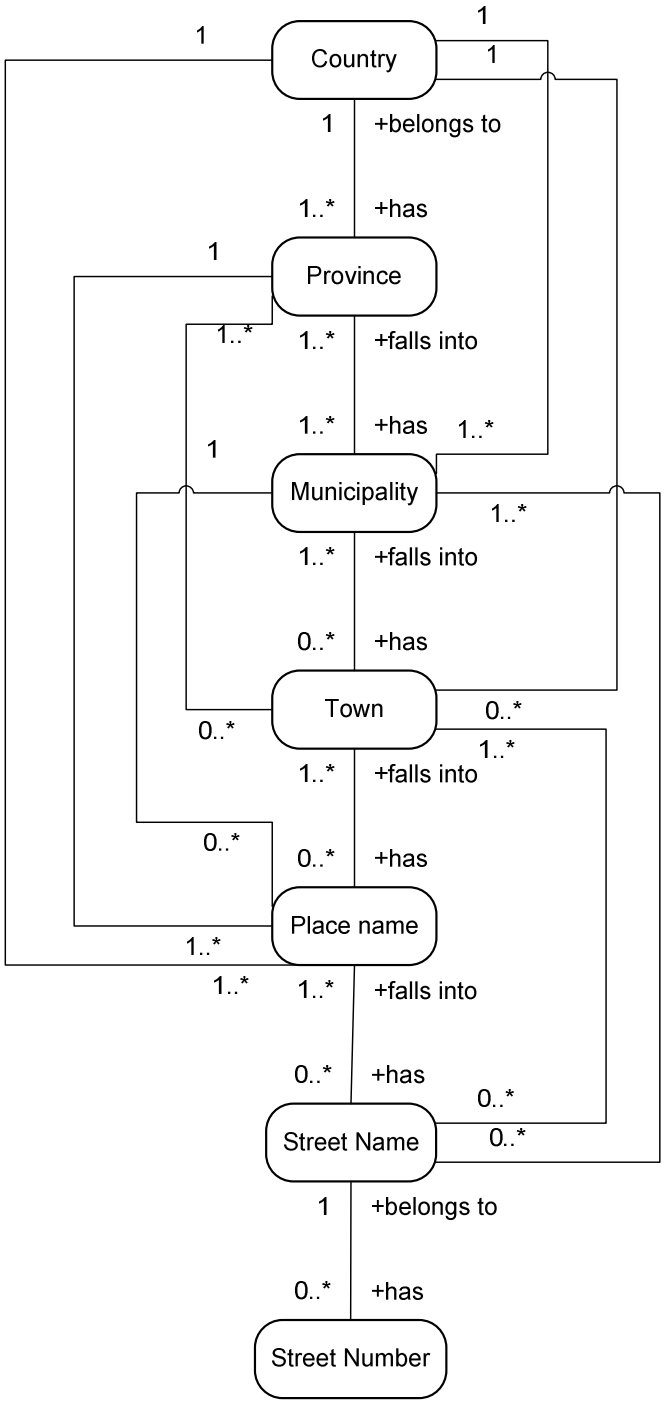


Figure 26. Relationships between the location types of the SANS 1883 street address type

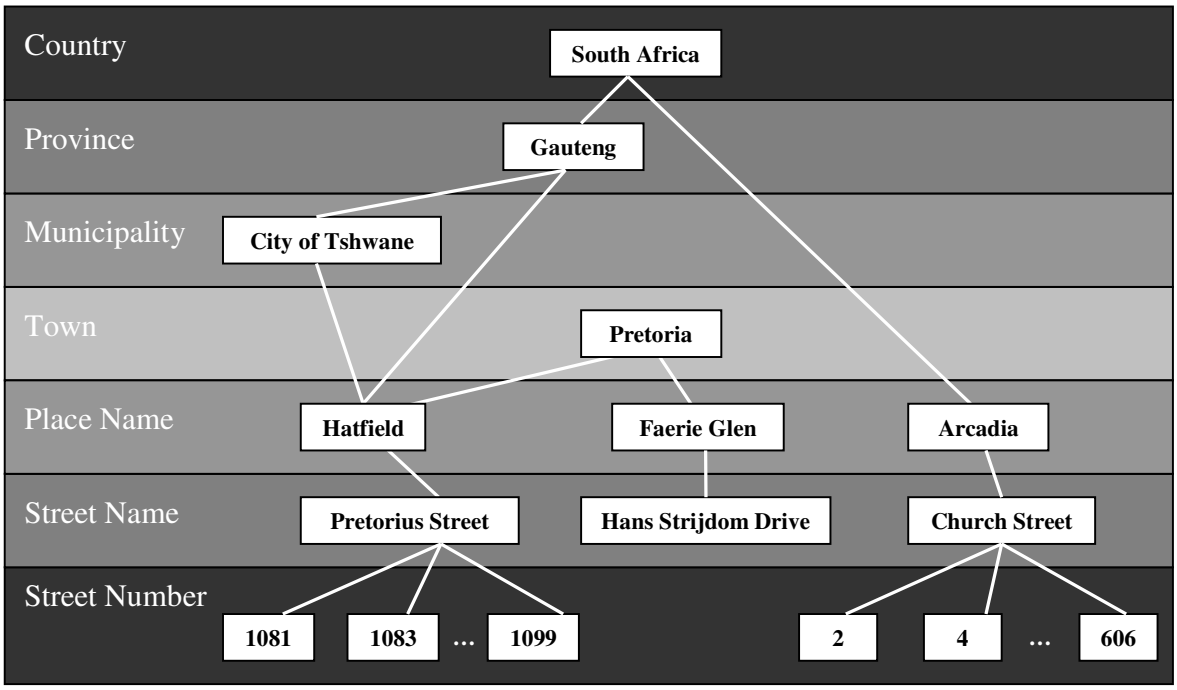


Figure 27. Instances of valid combinations of location types in the SANS 1883 street address type

4.3.2 Address data catalogue

The Compartimos catalogue contains information about addressing systems, address datasets and their associated access services and providers, node hosts, and address-related services and their providers. Thus, the catalogue includes all the metadata that is required for the operation of the address data grid. Figure 28 shows the relationships between the different elements of the catalogue. For simplification reasons, the data usage and data update notifications have been omitted from the model. The catalogue contains four collections: one of addressing systems, one of address dataset publications, one of address-related services, and one of node hosts. The addressing systems describe the types of addresses that are contained in an address dataset. A dataset can have addresses of more than one type, e.g. street addresses and intersection addresses. A dataset is published on the address data grid by associating it with an address data access service. Information about where a dataset is replicated is also stored in the catalogue. An address service provider provides address-related services that operate on the single virtual address dataset, such as for example, geocoding or mapping. The node host provides the resources to host some or all of the catalogue, replica, transfer and virtual address data service, as described in the computational and engineering viewpoint.

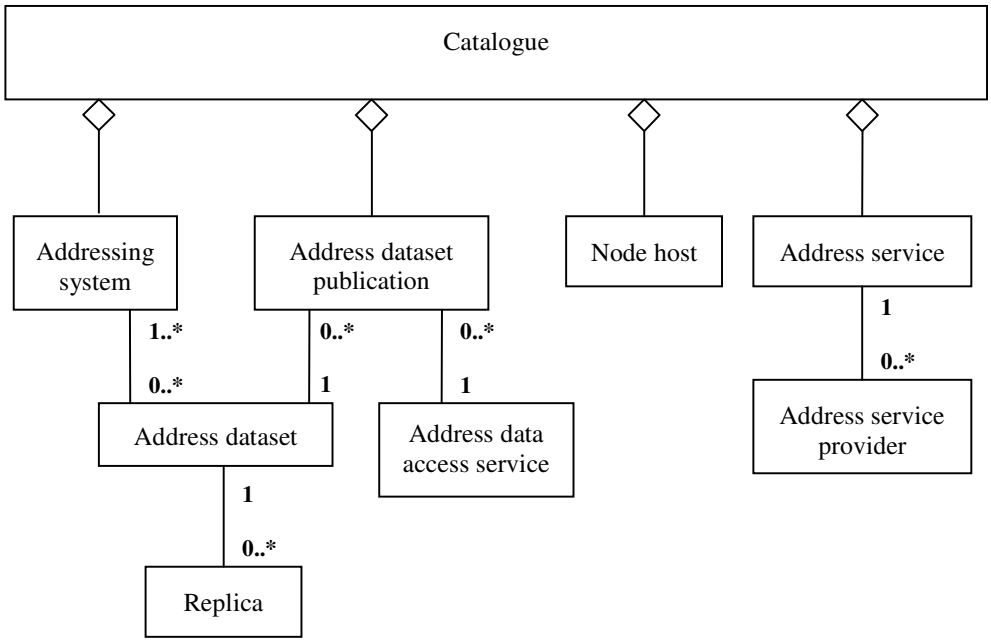


Figure 28. The address data catalogue

Data types and classes from ISO 19115 are used in the catalogue data model of Compartimos. This implies that existing ISO 19115 compatible metadata for address data can be imported into the catalogue for the address data grid. The information about the dataset includes descriptive information as defined in ISO 19115, as well as digital rights information, i.e. who may access which parts of the data and when. The details of the catalogue data model are included in Figure 55 of Appendix B. Examples of catalogue data is included in section B.3 of Appendix B.

Each information element in the catalogue has a status flag that is used, for example, to temporarily disable a dataset publication in the data grid without having to delete its information and later having to add it back again.

4.4 Computational viewpoint

In this section the *computational viewpoint of* Compartimos, which is concerned with the functional decomposition of the address data grid into a set of objects that interact at interfaces, enabling the single virtual address dataset. The OGSA data architecture describes the interfaces, behaviors and bindings for manipulating data within the broader OGSA, and Compartimos is based on this architecture. This implies that Compartimos follows a service-oriented approach similar to OGSA, and is in line with the kinds of services that are proposed in the OGSA data architecture. Where applicable, the details of these services in the OGSA data architecture are filled in to make provision for address data in an SDI environment. Compartimos thus is a domain-specific

application of the OGSA data architecture, which could also be referred to as a 'profile' of the OGSA data architecture for address data in an SDI.

The subsequent sub-sections first provide an overview of the different Compartimos objects and then go on to describe the purpose and capabilities of each individual object. For each object, its relation to the OGSA data architecture is discussed. Finally, in section 4.4.10 sequence diagrams of the use cases presented earlier in the enterprise viewpoint (refer to section 4.2) show when and why Compartimos objects interact with each other during these use cases.

4.4.1 Object overview

Compartimos comprises the following objects:

- the catalogue service (CatalogueService);
- the catalogue (Catalogue);
- the virtual address data service (VirtualAddressDataService);
- the address data access service (AddressDataAccessService);
- the data replica service (ReplicaService);
- the data transfer service (TransferService);
- the address dataset (AddressDataset); and
- the address-related service (AddressService).

The word 'object' is used here in compliance with the RM-ODP where it is used in the broader sense of the word and not with its very specific interpretation in the object-oriented paradigm. Table 11 provides an overview of the objects while Figure 29 shows how the Compartimos objects interact with each other in the address data grid. The Consumer, Address data provider and Address service provider objects are external to Compartimos and are therefore shown in a different color.

The hosting of objects is illustrated in Figure 34 of section 4.5, the engineering viewpoint of Compartimos, where it is discussed in more detail later. In this section, the computational viewpoint of Compartimos is discussed by describing objects and their interactions with other objects. While the replica and transfer services are more generic in nature and adopted in Compartimos from the OGSA data architecture with few or no modifications, other Compartimos services are tailored specifically for address data in an SDI environment. Some aspects of the OGSA data architecture, such as policies, storage management and caching, are excluded from Compartimos because they can be used generically for any kind of data and do not have to be tailored specifically for address data.

Table 11. Overview of the objects in the reference model

Object name	Type	Main purpose
CatalogueService	Service	Provides read and update access to the catalogue
Catalogue	Data	Stores information about services and data
VirtualAddressDataService	Service	Consolidates data
AddressDataAccessService	Service	Provides uniform access to individual address datasets
ReplicaService	Service	Replicates data in the address data grid
TransferService	Service	Transfers large volumes of address data
AddressDataset	Data	The individual address data set
AddressService	Service	A third party address-related service such as routing or mapping

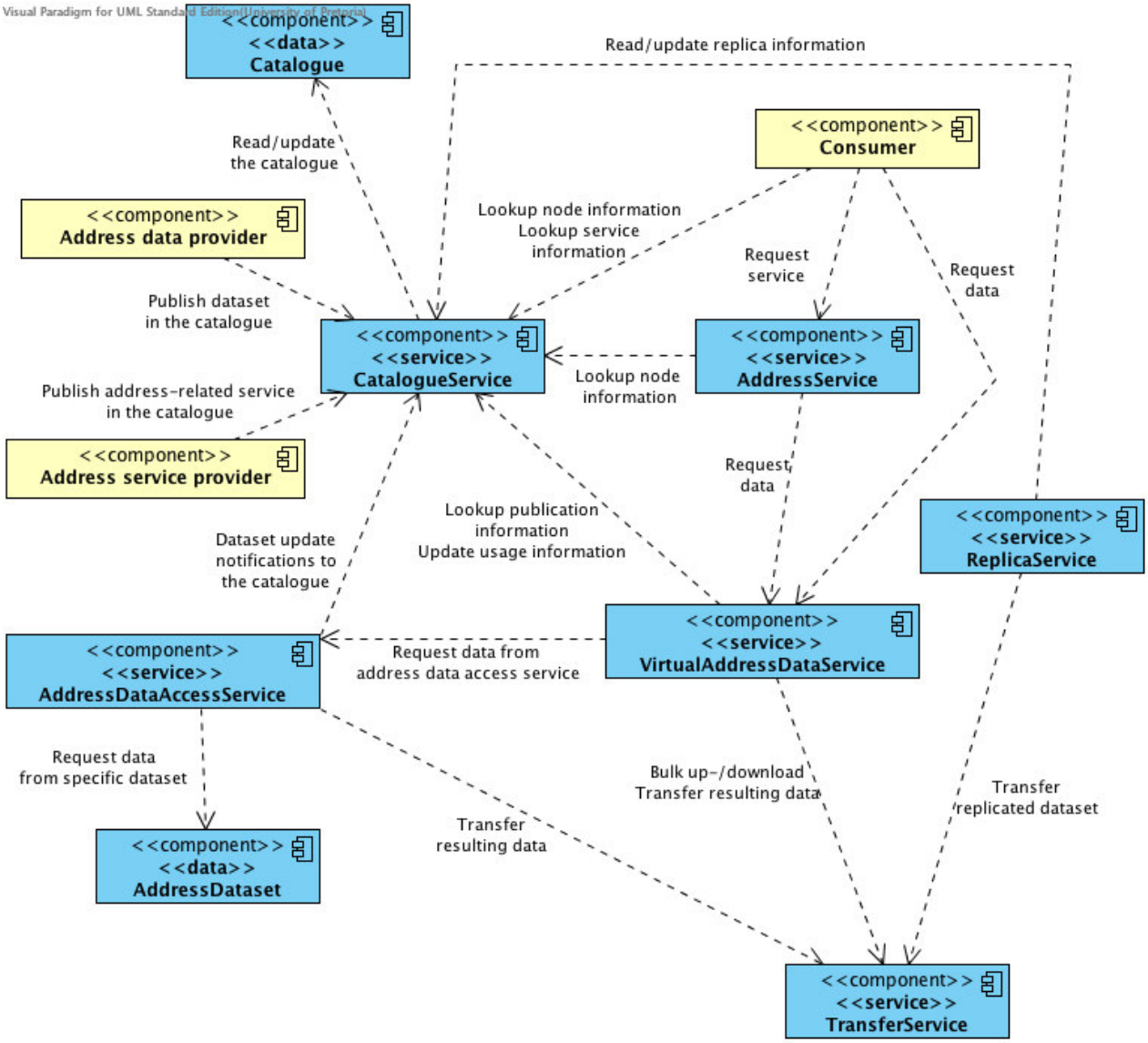


Figure 29. Object interaction in Compartimos

4.4.2 The catalogue service (CatalogueService)

The main purpose of the catalogue service is to provide read and write access to the information that is stored in the Compartimos catalogue. Inline with the OGSA data architecture, the Compartimos catalogue service provides *Publish* (add an entry), *Update* (modify an existing entry), and *Find* (apply query and return matching entries) services. The *Augment* (add additional properties for an entry created by someone else), *AddClassification* (add classification scheme) and *Classify* (classify an entry) services from the OGSA data architecture are not included in Compartimos. Since Compartimos applies to a very specific kind of data, these three services are not required. However, if in future, Compartimos is revised to include any kind of spatial data (traffic lights, road network, etc.), these three services will be relevant again. Table 40 in Appendix C provides a detailed list of all the operations that are provided by the catalogue service.

Addressing systems can be linked to more than one dataset; therefore modifications to an addressing system have to be coordinated among the providers of the relevant datasets. As discussed in the information viewpoint, it is expected that these addressing systems will represent the national address standards of different countries and therefore it is not expected that these addressing systems will change frequently. The version number attribute of the addressing system allows one to distinguish between different versions of the same addressing system, allowing co-existence and migration from one version of an addressing system to another.

4.4.3 The catalogue (Catalogue)

The Compartimos Catalogue object refers to the catalogue that was described in the information viewpoint in the earlier section 4.3.2. Any interaction with the catalogue takes place through the CatalogueService interface.

4.4.4 The replica service (ReplicaService)

The ReplicaService is responsible for replicating address datasets for fault tolerance, faster access and for scalability reasons. Replicas of datasets are stored on additional storage that is provided at the different node hosts. A node opts to allow replication or not. Datasets are either replicated as a whole, or parts thereof. There are different ways of splitting up a dataset for replication, for example, by selecting a geographic region of the dataset, by selecting specific address types, or by selecting addresses based on their creation date. An alternative way of splitting up an address dataset is to replicate the values of higher-level location types, thus providing an index into the dataset and speeding up, for example, an iterative data request. As an example, a street address such as *1083 Pretorius Street, Hatfield, Pretoria, South Africa*, that is captured on a user interface where the user is first presented with a combo box to select an addressing system, then a country, a province, a town and so forth. Due to the hierarchical nature of addresses, the higher

levels of location types such as the country, province, municipality and town contain far fewer instances than the lower levels. Therefore it makes sense to replicate the higher-level location types at as many nodes as possible, in order to speed up turnaround times for requests for this data.

In Compartimos information about the replica, i.e. the location, what is replicated, etc., is stored in the data catalogue and this information is accessible through the operations of the catalogue service. The ReplicaService is responsible for creating, deleting, validating, modifying the contents, and synchronizing the replicas of a dataset, however, in close coordination with the catalogue service: the ReplicaService updates the CatalogueService with information about the replicas whenever necessary. In turn, the VirtualAddressDataService discovers replicas through the catalogue. A dataset is replicated only if its data provider allows this by setting the appropriate attributes upon registration of the dataset in the catalogue, and the security policies of the original dataset have to be maintained by the replicas. Details of the operations provided by the ReplicaService are available in Table 41 of Appendix C.

The ReplicaService implements the replication strategy, i.e. *when* a dataset is replicated to *where*. The VirtualAddressDataService updates data usage information in the catalogue, which the ReplicaService reads and uses to implement the replication strategy. Compartimos does not prescribe a specific replication strategy so that different replication strategies or variations thereof can be employed in the address data grid over a period of time, depending on the current circumstances. The Compartimos approach, similar to the OGSA data architecture, isolates the ReplicaService as an object on its own, and provides a well-defined interface for the ReplicaService which brings the advantage that the ReplicaService can be exchanged over time: a plug-and-play approach, so to speak.

4.4.5 The transfer service (TransferService)

The TransferService moves data between node hosts, data hosts, and data consumers. This data movement could be the result of a data request, or the result of dataset replication being required. The TransferService is used by the ReplicaService for replicating data, by the VirtualAddressDataService for transferring large data results and for uploading address data in bulk. Note that requests for data will not always have to make use of the TransferService. It is only required when the resulting dataset is large, such as, for example, a request for address data for the whole of the Gauteng province in South Africa. In line with the OGSA data architecture, the Compartimos TransferService is protocol agnostic (i.e. supports various transport protocols as appropriate) and employs a lower level transfer protocol, such as GridFTP, to transfer address data in bulk from one location to another.

This service does not require customization or specialization for address data, and in

Compartimos mostly the same operations as in the OGSA data architecture are included: *SetupTransfer*, *PauseTransfer*, *ResumeTransfer*, *and StopTransfer*. The *CreateTransfer* service in the OGSA data architecture has been renamed to *StartTransfer* in Compartimos, and a *GetTransferState* operation, with which the state of the transfer can be monitored, as recommended by the OGSA data architecture, has been added. The details of the operations provided by the TransferService are listed in Table 42 of Appendix C. Similar to the ReplicaService, the TransferService is isolated as an object on its own, both conceptually as well as on implementation level, allowing the address data grid to employ different transfer services over a period of time.

4.4.6 The address data access service (AddressDataAccessService)

The AddressDataAccessService converts the address dataset from local proprietary format to the address data model described in the information viewpoint, acting as an interpreter for a specific source address dataset and providing a uniform access method to any dataset that is published in the address data grid. Thus, this service performs a role similar to that of an Open Database Connectivity (ODBC) driver, a vendor-neutral, standardized, application programming interface (API) for accessing SQL databases. The AddressDataAccessService also has the responsibility to notify the catalogue of updates in the datasets associated with it so that replicated datasets can be synchronized, when necessary.

The OGSA data architecture proposes three generic data access operations for structured data: *Create*, *ExecuteQuery* and *BulkLoad*. The *Create* operation creates an association between a data service and an underlying data resource, which may be created and populated as a result of this operation. In an SDI environment, the main drive for an address data grid is to publish existing address data that is maintained locally; therefore this operation has been adjusted slightly for use in Compartimos by providing a *CreateDataset* operation with the AddressDataAccessService and a *RegisterDataPublication* operation with the CatalogueService.

The *RegisterDataPublication* associates a dataset with an *AddressDataAccessService*. The Compartimos model provides for a one-to-many relationship between a dataset and an access service, allowing more than one access service to be associated with the same dataset and thereby increasing scalability, i.e. while the single dataset still has to execute the raw queries in series, translation into the interoperable Compartimos address data model can be done in parallel. Multiple data access services per dataset also enable versioning of the address data model in the Compartimos catalogue: each service can support a different version of the address data model.

In Compartimos the ReplicaService uses the *CreateDataset* operation of the AddressDataAccessService to create a replica. Once this replica of an original dataset has been created and populated, its information is added to the catalogue, and it can be used in subsequent

data queries. Thus, in Compartimos the physical creation of the dataset is separated from adding the association between an address dataset and an address data access service to the catalogue. This separation is reflected in the *1..0** relationships between an address dataset publication and its associated dataset and address data access service in Figure 28.

The *ExecuteQuery* operation is represented by the *GetAddress* operation of the AddressDataAccessService and the *BulkLoad* operation is represented by the *UploadAddressData* operation in Compartimos, performing more or less the same functionality as in the OGSA data architecture, albeit customized for address data. Details of the operations of the AddressDataAccessService can be found in Table 43 of Appendix C.

4.4.7 The virtual address data service (VirtualAddressDataService)

The VirtualAddressDataService provides the required consolidation functionality to make the distributed heterogeneous address datasets appear to be a single virtual address dataset. The VirtualAddressDataService uses the CatalogueService to discover datasets and/or their replicas that could satisfy an incoming request for data.

Any incoming data request or data query specifies its requirements in terms of data currency. For example, for a general mapping application it is sufficient to return address data from a dataset that was replicated a week ago and has been updated in the mean time, but an address data request for authentication by a financial institution requires the latest version of the dataset and should force synchronization before returning the results. While the AddressDataAccessService interprets proprietary address data formats and converts them to the interoperable Compartimos address data model described in the information viewpoint, the VirtualAddressDataService is responsible for all other consolidation, such as removing duplicates (resulting from the same address occurring in multiple address data sources) and resolving ambiguities. This is also the service where address-related intelligence, such as matching incomplete addresses that are supplied as filter of a *GetAddress* operation, are matched to addresses requested from individual data resources.

The OGSA data architecture defines a set of operations for a Data Federation service, which is defined as the logical integration of multiple data services or resources so that they can be accessed as if they were a single data service. In a way this corresponds to the VirtualAddressDataService in Compartimos, however, OGSA operations provide the functionality to associate a number of resources into a single federation. Example operations are *CreateFederation*, *AddSourceToFederation*, *AddAccessMechanism*, and *UpdateFederationAttributes* and a wide variety of services ranging from input data resources to transformations of data and filters can be federated. In Compartimos a dataset (the resource) is automatically included in the federation when it is published in the catalogue and resources are, by definition, limited to address datasets. Therefore,

Compartimos provides only for the *GetAddress* and *UploadAddressData* operations, which mirror the AddressDataAccessService operations with the same name. The main goal in an SDI environment is to publish address data and therefore the *CreateDataset* and *AddAddress* operations, which are part of the AddressDataAccessService for replication purposes, are not required on the level of the VirtualAddressDataService. Details of the VirtualAddressDataService operations can be found in Table 44 of Appendix C.

4.4.8 The address dataset (AddressDataset)

The Compartimos AddressDataset object refers to any address dataset that is published on the address data grid. In OGSA data architecture terminology this is the data source or data resource. The *RegisterDataPublication* operation of the catalogue service associates an address data access service with a particular address dataset, and from then on the AddressDataset is available for inclusion in address data queries and requests on the grid. While the particulars of the underlying dataset, such as the format, data model, etc., influence the performance of data access, they are not important in Compartimos since the AddressDataAccessService provides the interpretation to the interoperable Compartimos address data model.

4.4.9 The address-related service (AddressService)

The AddressService refers to any address-related service, such as routing or mapping, that is offered by a third party on top of the single virtual address dataset in the grid. The list of operations of the address-related service is application dependant and defined by the service provider. The AddressService interacts with the VirtualAddressDataService when executing its address-related service.

4.4.10 Object interaction

This section describes and illustrates interaction of the Compartimos objects during the three use cases that were presented earlier in the enterprise viewpoint of section 4.2. UML sequence diagrams are used to illustrate these interactions that effectively realize the address data grid in an SDI. Four additional sequence diagrams for uploading an address dataset, publishing an address dataset on the grid, publishing an address-related service on the grid, and dataset replication are provided in Appendix D.

4.4.10.1 Simple data request

In the simple data request the user specifies a filter such as a bounding box, and the address data grid returns all the data that is available within this bounding box. For example, a mapping application could request addresses that are to be displayed on its map. The mapping application interacts with the VirtualAddressDataService only. The VirtualAddressDataService handles the

execution of the distributed query by finding relevant datasets, or replicas of them, and their associated data access services through the CatalogueService, and then requesting the specified data from the various address datasets through the respective AddressDataAccessServices. The VirtualAddressDataService consolidates the resulting data by, for example, removing duplicates addresses, and then returns the consolidated resulting dataset to the mapping application. Figure 30 shows how the data is returned as parameters of a service request, while Figure 31 shows the object interaction when the resulting data is returned as a file through the TransferService.

4.4.10.2 Iterative data request

In the iterative data request (refer to Figure 32) the level of detail of the requested address data is increased iteratively with each subsequent request. This type of request is required to allow a user to select an address from dropdowns, for example, when capturing their residential address the dropdowns guide customers in selecting an address that is valid by first presenting a list of addressing systems, then a list of, for example, provinces, municipalities, suburbs, streets and so on. For simplicity reasons, the sequence diagrams below include interaction with a single address dataset, but the location type values could be requested from more than one dataset and consolidated by the VirtualAddressDataService. The diagrams show the case where location type values are returned as parameters in a service request, but similar to the simple data request, the TransferService can be used to transfer the resulting data in a file.

4.4.10.3 Service request

In the sequence diagram for the service request use case (refer to Figure 33) the customer interacts with the VirtualAddressDataService only. The sequence diagram does not show the details of the object interaction for the simple data request, which are illustrated in Figure 30. Note that the AddressService could also invoke an iterative data request. The diagram shows the case where service results are returned as output parameters, but if necessary, they could also be returned via the TransferService.

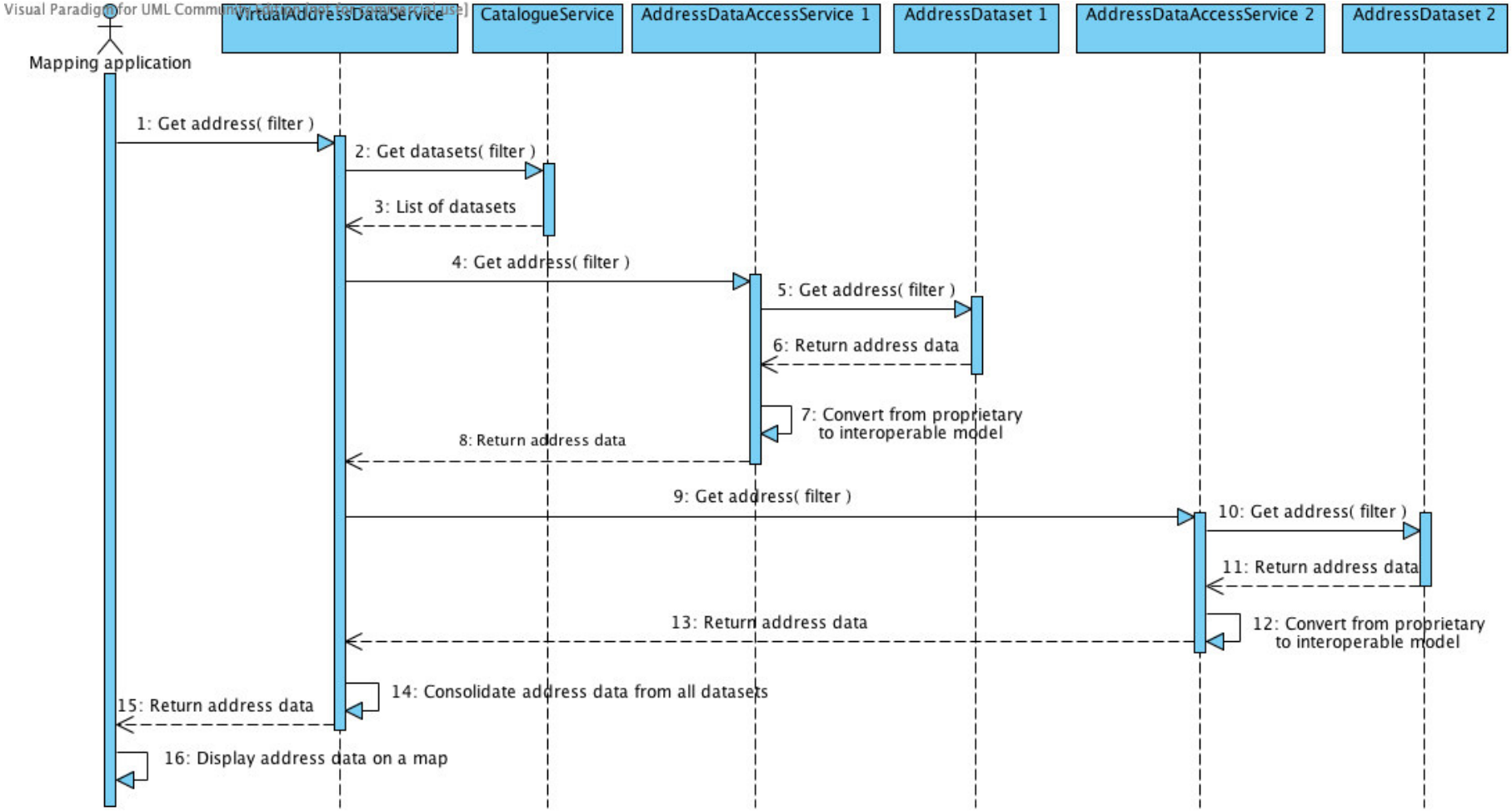


Figure 30. Simple data request (sequence diagram)

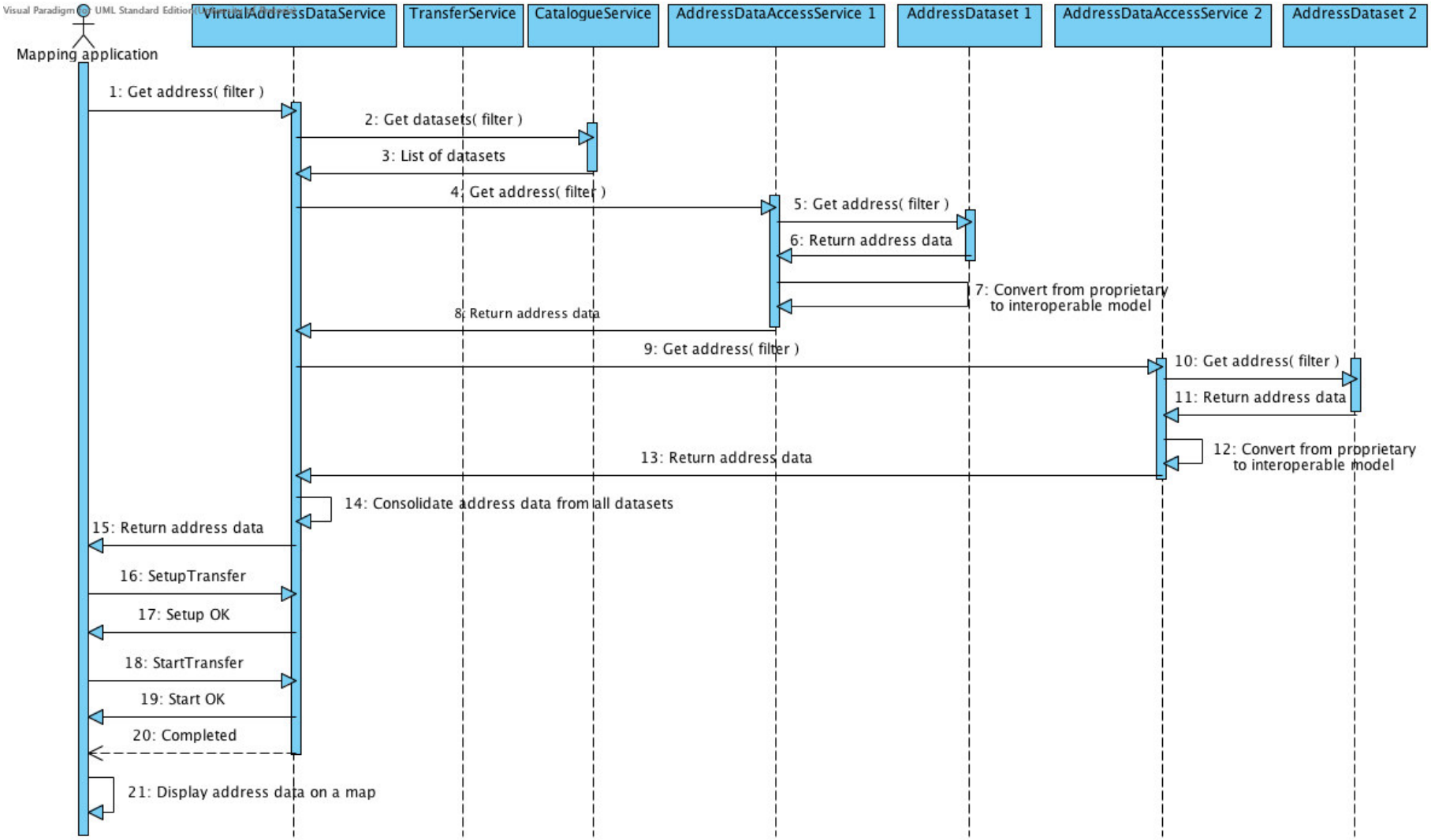


Figure 31. Simple data request involving the TransferService (sequence diagram)

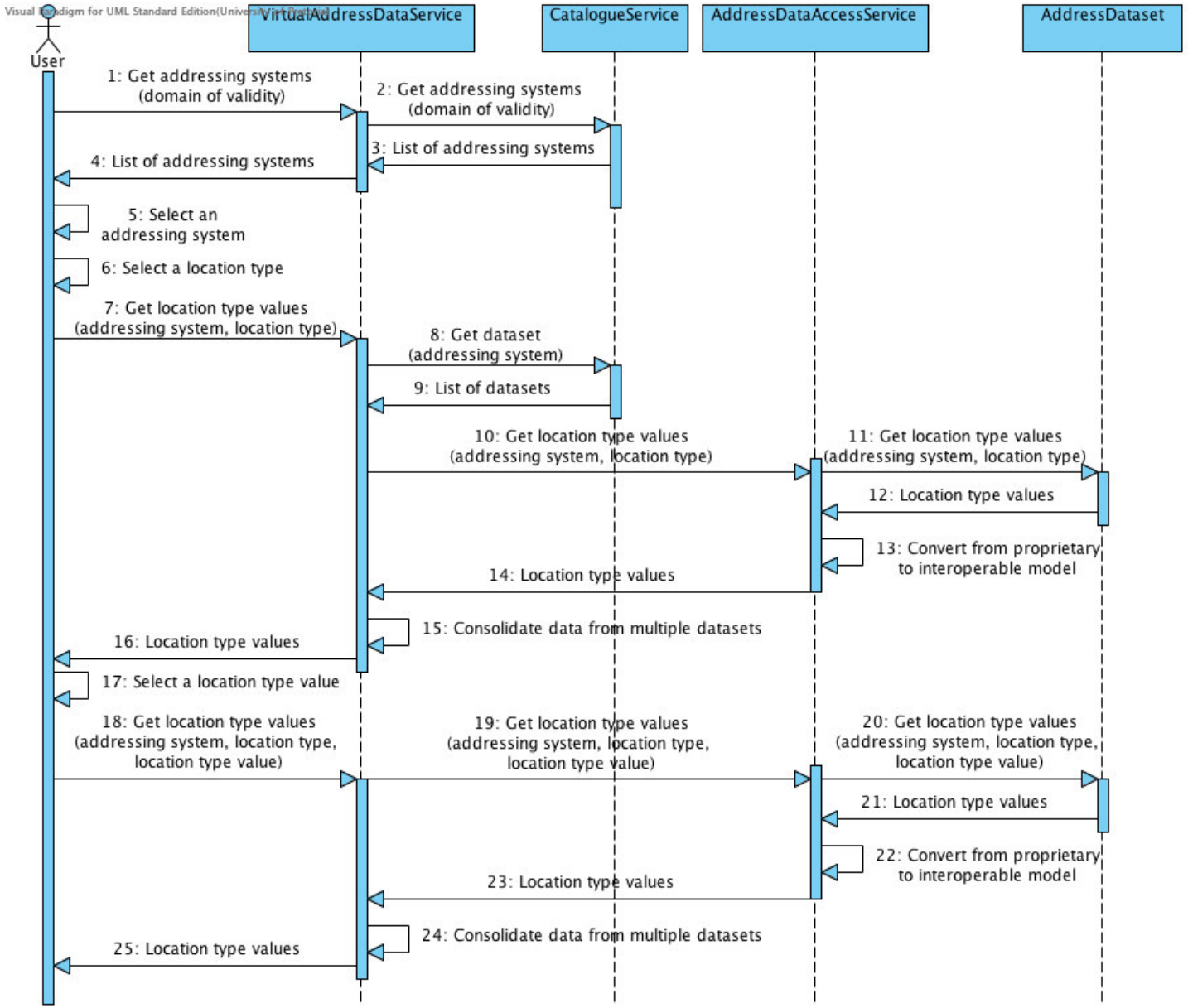


Figure 32. Iterative data request (sequence diagram)

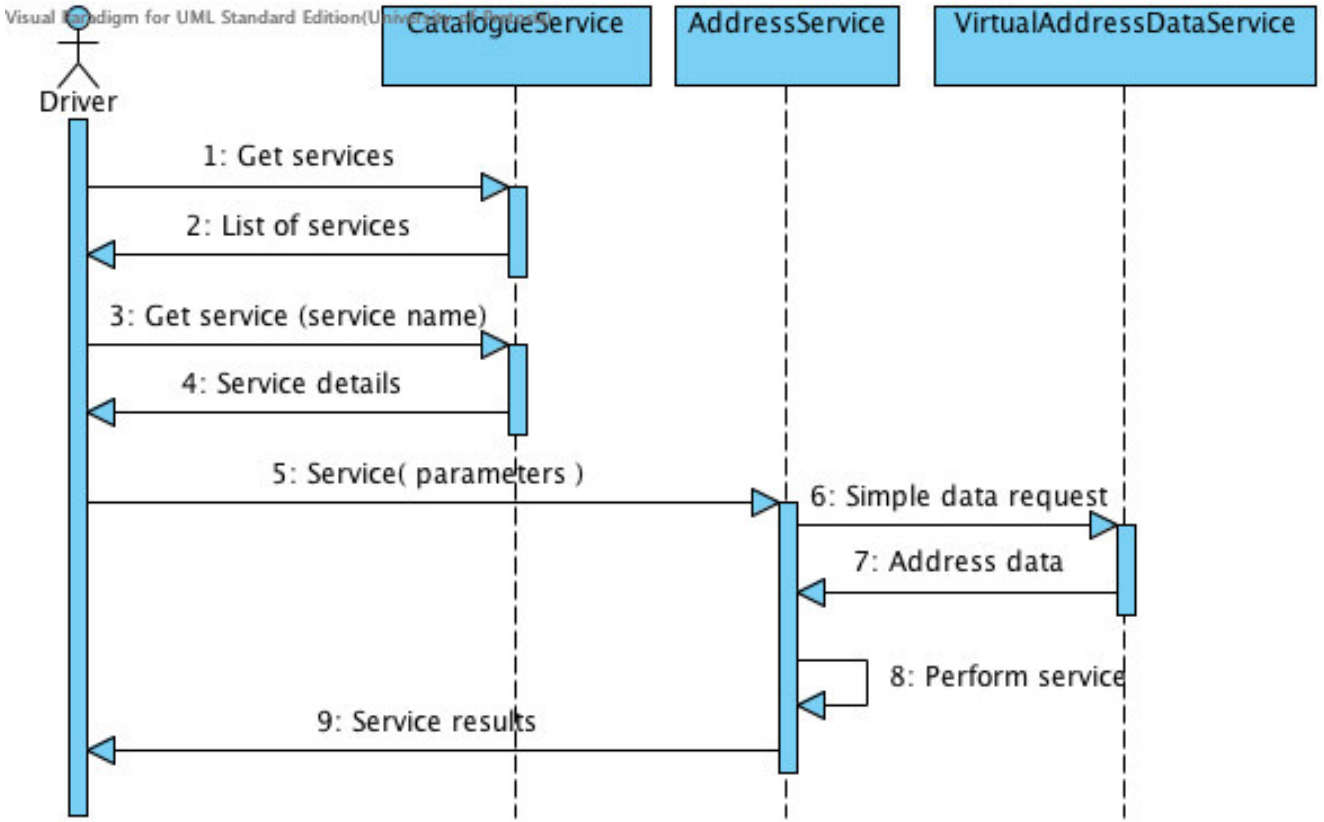


Figure 33. Service request (sequence diagram)

4.5 Engineering viewpoint

In this section the *engineering viewpoint of* Compartimos is presented, which is concerned with the infrastructure required to support the virtual address dataset. While the computational viewpoint describes when and why objects interact, the engineering viewpoint describes how objects interact and which resources are required for this interaction. Thus, in this viewpoint details about potential deployments of the Compartimos objects are included.

4.5.1 Object deployment

In Compartimos, there are three types of hosts. Firstly, at the data host the dataset and the AddressDataAccessService are hosted. Secondly, at the node host the CatalogueService, VirtualAddressDataService and the TransferService are hosted, and optionally also a ReplicaService. The master catalogue is located at one of the node hosts; which one it is hosted at, depends on the replication strategy, which is explained further below. Thirdly, at the service host the address-related AddressService is hosted. Figure 34 shows the three types of hosts with the objects deployed at each.

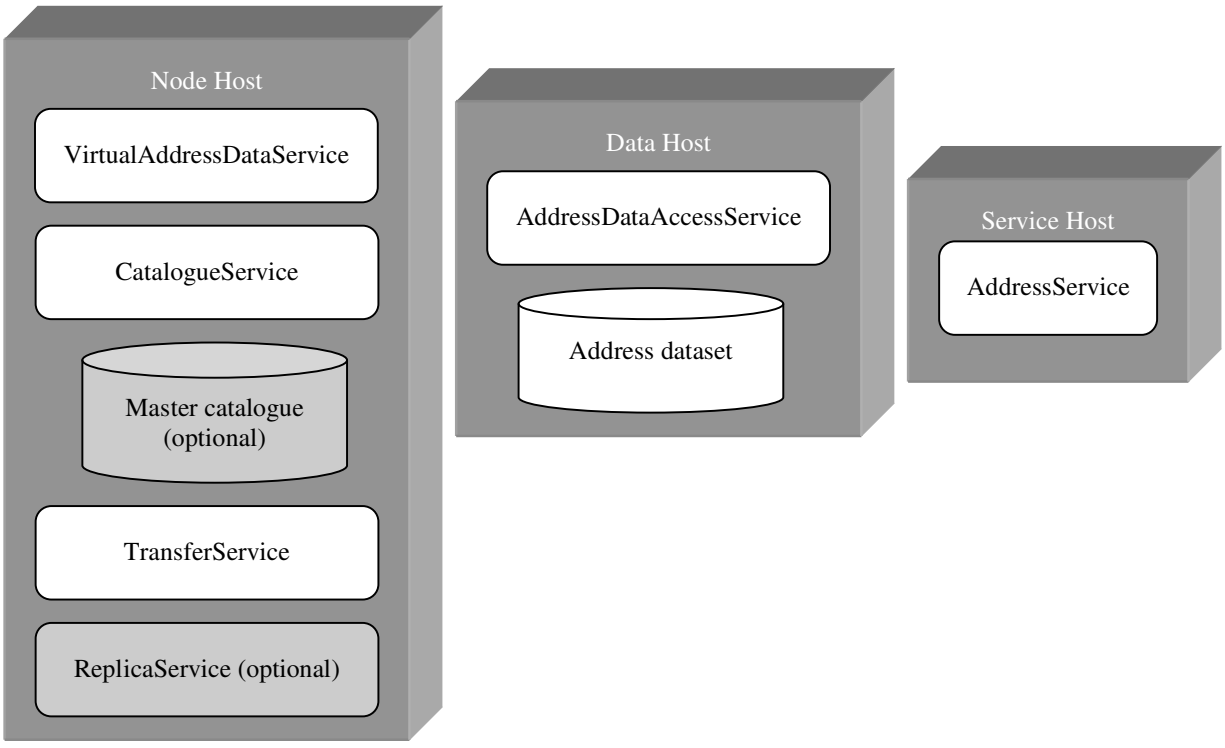


Figure 34. Three types of hosts in Compartimos

Note that the ReplicaService and master catalogue are optional at the node host. The ReplicaService is only required if the node host opts to provide additional storage space for dataset replication. Figure 34 shows that the master catalogue has an optional location at the node host. Catalogue information can be split into three types of information: firstly, information about data and services, secondly information about replicas, and thirdly, information about data usage. There is only one master copy of the catalogue in Compartimos, which is always located at a node host (the data and service host are restricted to providing data and providing services). In an SDI environment, datasets are not continuously published. Rather, data providers publish their datasets once, and only when the underlying proprietary structure of the address dataset changes so that it warrants a new AddressDataAccessService, does it become necessary to update the catalogue. In other words, the information about data and services in Compartimos is relatively static. Replica information will change more frequently, but still not on a daily basis. Data usage information changes frequently, but this information could be cached locally at a CatalogueService and updated at certain time intervals. Therefore, a rather simple replication strategy can be employed for the updating of the master catalogue. Compartimos does not prescribe a replication strategy, but the following strategy could, for example, be used:

- Each local CatalogueService keeps a replica of the master catalogue, which it uses for queries.
- Updates to the catalogue are routed from the local CatalogueService to the master catalogue, from where they are propagated to all the replicas.

This strategy results in lots of network traffic when there are catalogue updates but it is expected that after an initial set-up period the updates to the information about data and services will ‘cool down’ and become minimal.

As explained earlier in the enterprise viewpoint of section 4.2, a single institution could be both a data and a node host. Figure 35 shows a potential deployment of Compartimos with a variety of hosts, each hosting a different combination of Compartimos objects.

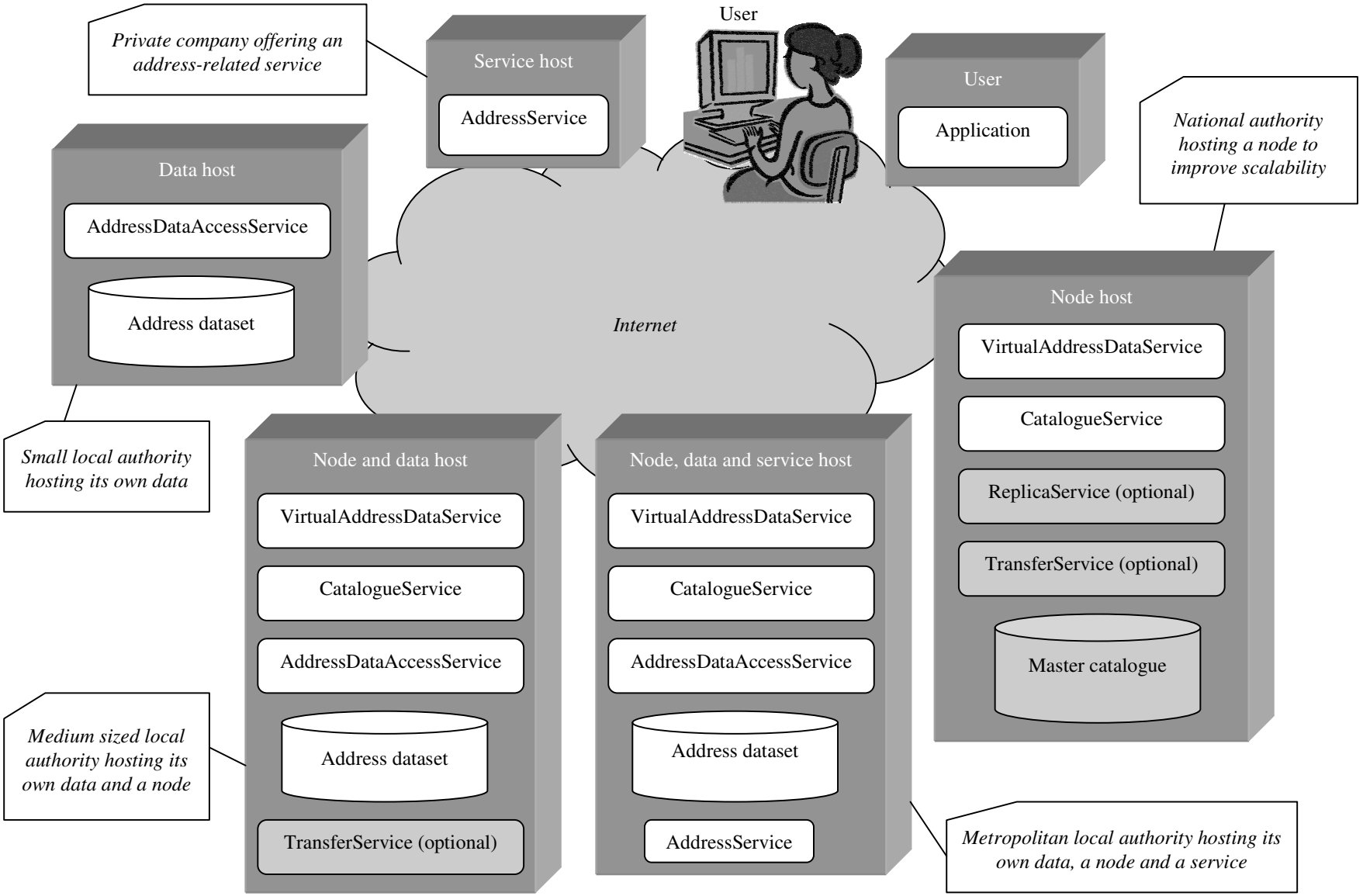


Figure 35. Deployment diagram for the address data grid with a variety of hosts

4.6 Discussion

In this section Compartimos is related to the OGSA data architecture, of which it is a special case or profile. The first subsection provides a general comparison, while the second and third subsections relate the concepts of virtualization and service-orientation as they apply to Compartimos.

4.6.1 Comparison overview to the OGSA data architecture

The OGSA data architecture presents a "toolkit" of data services and interfaces that can be composed in a variety of ways to address multiple scenarios. These services and interfaces include data access, data transfer, storage management, data replication, data caching, and data federation (OGF 2007a). The components of the data architecture can be put together to build a wide variety of solutions and Compartimos is one example of such a solution. Compartimos gives an abstract representation of the essential components of an address data grid in an SDI and is a profile or specialization of the OGSA data architecture, illustrating how a specific solution based on the OGSA data architecture can be designed. While Compartimos focuses on address data, it serves as an example for other kinds of spatial data, such as points of interest, traffic lights or manholes that are also produced and maintained by individual local authorities. Compartimos is one of the first attempts in the joint SDI and data grid problem space, and thus enhances the mutual understanding of these two domains.

Table 12. Services in the OGSA data architecture and related services in Compartimos

OGSA data architecture	Compartimos
Data Transfer	TransferService
Data Access	AddressDataAccessService
Storage Management	Not included in Compartimos*
Cache Services	Not included in Compartimos*
Data Replication	ReplicaService
Data Federation	VirtualAddressDataService (federation *and* consolidation, the latter is not included in the OGSA data architecture)
Data Catalogues and Registries	CatalogueService

* No need for specialization. Generic grid-enabled services are sufficient.

Similar to the OGSA data architecture, Compartimos follows a service-oriented approach and the OGSA data architecture services are specialized to make provision for address data in an SDI environment in Compartimos. Compartimos also includes an interoperable address data model, and

the catalogue information is based on the ISO 19115 standard has been adopted in SDIs around the world. The address data model is based on ISO 19112, another standard in the ISO 19100 series of standards. Table 12 provides a summary overview of the services in the OGSA data architecture and their counterparts in Compartimos.

The *Augment* (add additional properties for an entry created by someone else), *AddClassification* (add classification scheme) and *Classify* (classify an entry) services from the OGSA data architecture are not included in Compartimos. Compartimos applies to a very specific kind of data, and therefore these services are not required. However, for a reference model that accommodates any kind of geographic data, such as applicable within ISO/TC 211, these services will be relevant and should be included.

The OGSA data architecture includes a section on security that describes issues that are important in a data grid. Specific security-related services are not included but it is recommended that all services should:

- Advertise the degree to which they adhere to security requirements.
- Accept security related information in their interfaces.
- Pass security related information, such as security credentials in all service requests from this service. This security information may be held within the service or may have been provided as part of an invocation of this service.

Similarly, to ensure data privacy, the following issues need to be addressed by all services in the OGSA data architecture:

- The set of access requests from a user may need to be private to that user. This impacts the logging of those queries by the data service.
- Privacy of data needs to be assured when at rest (e.g., on disk or tape). This may require encryption of data when it is at rest.
- Privacy of data in transit (e.g., the result of a data access request) must be ensured. This may require encryption in the communication channel.
- A data service should advertise the degree of privacy that it supports.

Regarding security, all of the above are also applicable in an address data grid in an SDI. The Compartimos address data model deliberately excludes any information about the person(s) or business residing at an address, to protect their privacy. One other aspect that is worthwhile mentioning in an SDI context is the questions of trust: which address data sources can the data grid trust to be accurate? In many countries a residential address is a prerequisite for opening a financial

account. If the address data grid is used for residential address verification, it is imperative that it is verified against legally valid addresses only. This can be achieved by making use of the metadata associated with an address to include only address data from custodians in the address verification. In countries, such as South Africa, where custodians for address data have not been assigned, this will not work and one has to explore other mechanisms, such as calculating a confidence level for the address based on, for example, its occurrence in or omission from a number of address datasets. These mechanisms are not described in detail in Compartimos and there is room for future work on this topic.

4.6.2 The layered aspect of Compartimos

Compartimos follows the same layered approach that is applied in grids and that allows virtualization, as described earlier in Chapter 3 . Figure 36 illustrates the layers presented in 3.3.1 to show where the Compartimos services (in bold italics) fit into that layered architecture.

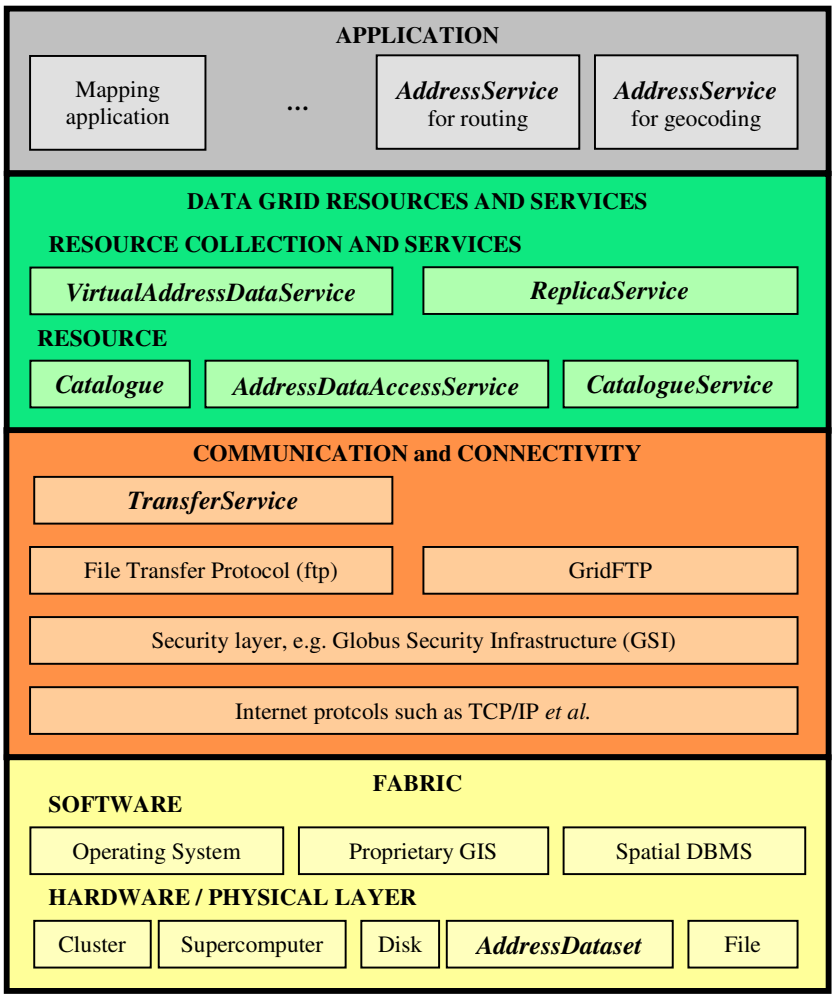


Figure 36. The Compartimos services in the four main layers of the Grid architecture

The distributed heterogeneous *AddressDatasets* (data sources) on the fabric layer are abstracted by the *AddressDataAccessService* on the resource layer into data sources with a uniform interface. The *Catalogue* and *CatalogueService* on the resource layer assist in this abstraction and virtualization by providing information about resources. The *TransferService* (along with the TCP/IP and other protocols) on the communication layer provides for connectivity between the *AddressDataset* and the *AddressDataAccessService*. Finally, both the *ReplicaService* as well as the *VirtualAddressDataService* operate on a collection of *AddressDatasets* (resources), and an application at the highest-level requests an address without being concerned about the details of the underlying consolidations, communication protocols and physical devices.

It is interesting to note that there are similarities to the assignment of components to Grid layers reported by Wei *et al.* (2006): the transfer service (RFT) is on a lower level than the grid-enabled catalogue service and the replica management service, and underlying it all is the Globus Security Infrastructure (GSI).

4.6.3 Service-oriented architecture of Compartimos

Compartimos follows the same service-oriented approach that was presented in Chapter 3 earlier, and that is similar to OGSA. Services for address data access (*AddressDataAccessService*), address data coordination (*VirtualAddressDataService*) and third-party address-related services (*AddressService*) are registered in the Compartimos catalogue. These services can be discovered, and are then bound to perform the services. Figures 37-39 illustrate the concept of service-orientation in Compartimos for address data access, nodes, and address related services, respectively.

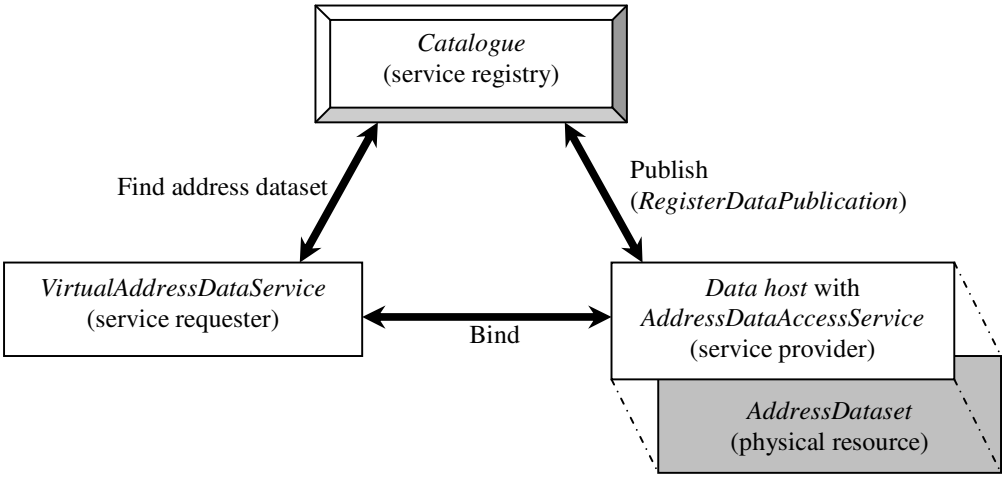


Figure 37. Service-orientation for address data access in Compartimos

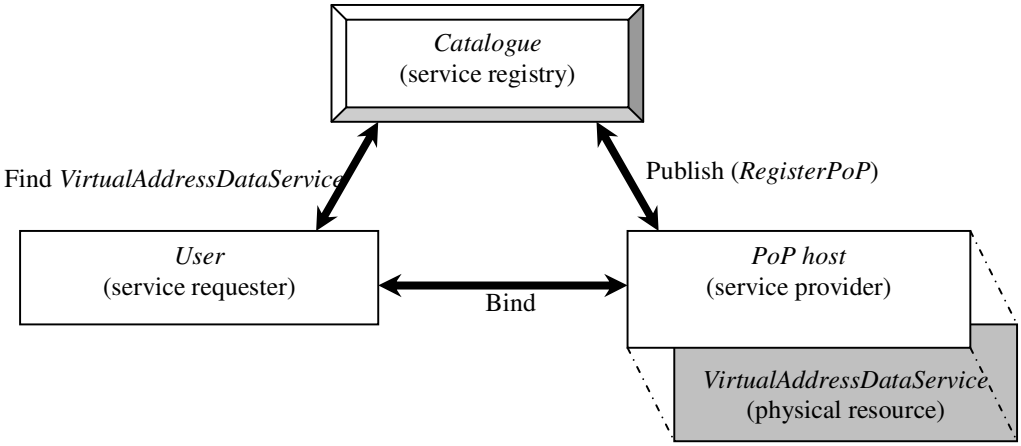


Figure 38. Service-orientation for nodes in Compartimos

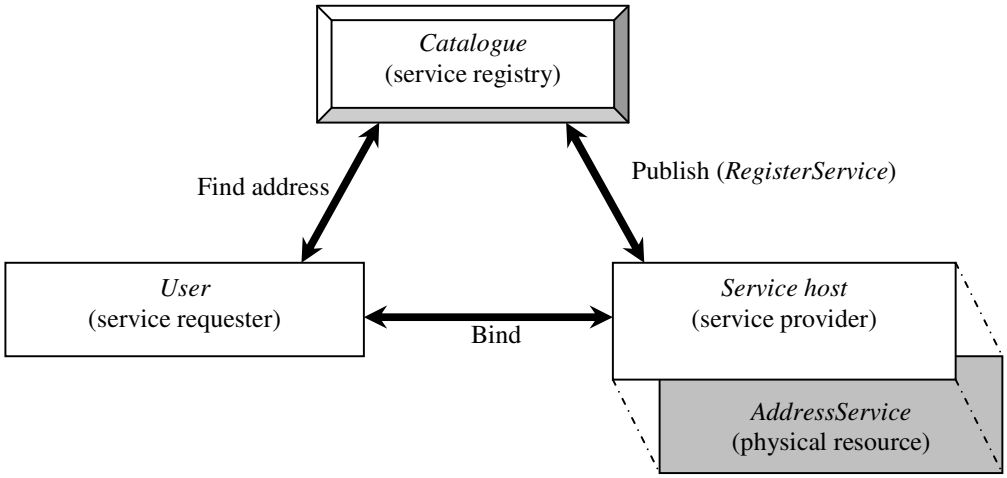


Figure 39. Service-orientation for address-related services in Compartimos

Chapter 5 Implementation and evaluation of Compartimos

5.1 Introduction

In this chapter the *technology viewpoint* of Compartimos is presented. This is the fifth and last of the RM-ODP viewpoints and is concerned with the choice of specific technologies in the implementation of Compartimos. The first section considers technology choices for specific Compartimos objects, while the second section of this chapter considers technology choices for Compartimos overall. Together these two sections (5.2 and 5.3) comprise the technology viewpoint, which contributes towards understanding what technologies are available to make an address data grid in an SDI a reality. The technology viewpoint presented here does not recommend specific technology choices, but rather discusses available choices. In section 5.4 the proof of concept implementation of Compartimos is described. This represents a very specific selection of technology choices. The novel evaluation framework that will be presented in Chapter 6, describes important criteria for a national address database in an SDI and in section 5.5 the Compartimos reference model is evaluated against these criteria. This evaluation contributes towards understanding the applicability of Compartimos for address data in an SDI. The chapter is concluded with a discussion of results and recommendations that are drawn from the experience of the design, implementation and evaluation of Compartimos. Some ideas for expansion on Compartimos are included.

5.2 Technology choices of specific Compartimos objects

Table 13 below provides an overview of available technology choices for Compartimos objects, which are discussed in further detail in the subsections of this section. These technology choices include technologies and standards from the geospatial as well as the grid community. The discussion is augmented with references to reports of where these technology choices have been implemented in related work. For each Compartimos object, there is also the option of developing it from scratch without using existing technology. As always when implementing something from scratch, this option has the advantage that there is no historic baggage that needs to be accommodated, but the disadvantage that it is more expensive due to learning curve and a larger amount of more work. The learning curve includes both learning what others have learnt before in their implementations, as well as users having to learn to use the newly developed object. These pros

and cons of software reuse have been well documented over the years (Tracz 1994, Morad and Kuflik 2005, Finnigan and Blanchette 2008) and this technology choice is therefore not specifically mentioned in the table and the subsequent discussion.

Table 13. Overview of technology choices for Compartimos objects

Compartimos object	Technology choices
Catalogue	Relational data model vs other models Relational DBMS vs other DBMS such as XML or Object DBMSs ISO 19112:2003, *Geographic information – Spatial referencing by geographic identifiers* ISO 19115:2003, *Geographic information – Metadata* ISO 19119:2003, *Geographic information – Services* ISO 19139:2003, *Geographic information – Metadata – XML schema implementation* Dublin Core metadata elements Metadata in the Monitoring and Discovery System (MDS) of the Globus Toolkit Metadata in the Replica Location Service (RLS) of the Globus Toolkit
CatalogueService	OpenGIS Catalogue Service Implementation Specification Monitoring and Discovery System (MDS) of the Globus Toolkit Replica Location Service (RLS) of the Globus Toolkit
ReplicaService	Data Replication Service (DRS) of the Globus Toolkit Replication capabilities available in DBMS and GIS software
TransferService	Reliable File Transfer (RFT) of the Globus Toolkit, making use of the File Transfer Protocol (ftp) and GridFTP protocol
AddressDataAccessService	OpenGIS Web Feature Service Implementation Specification (WFS) ISO 19142 (draft), *Geographic information – Web feature service* OGSA-DAI data resources, compatible with the Globus Toolkit Combination of OGSA-DAI data resources and WFS
VirtualAddressDataService	Intelligence: OGSA-DAI Distributed Query Processing (DQP) Address-specific toolkits such as AfriGIS Intiendo Any capabilities available as OpenGIS Web Processing Service Data: OGC Web Feature Service (WFS) ISO 19142 (draft), *Geographic information – Web feature service* OGSA-DAI data resources, compatible with the Globus Toolkit
AddressDataset	Technology independent, up to the data provider
AddressService	OpenGIS Web Feature Service (WFS) OpenGIS Web Processing Service (WPS)

5.2.1 The catalogue

The volumes of data in the catalogue are determined by the number of addressing systems, the number of dataset publications, the number of node hosts, and the number of registered address-

related services. It is difficult to estimate how many of these there will be in an address data grid. Ballpark figures, based on the largest possible scenario of an international address data grid where there is a national address dataset publication for each country, amount to hundreds of datasets (typically one or two per country) and maybe thousands of addressing systems (typically a few per country). Taking ISO 3166-1, *Codes for the representation of names of countries and their subdivisions - Part 1: Country codes* (2006) as a guideline, which includes approximately 250 country codes, and assuming that there are ten addressing systems in a country, this would amount to 2,500 addressing systems and 250 dataset publications registered in the catalogue. South Africa has twelve address types (SANS 1883 2008), but most other countries have less (AS/NZS 4819:2003, Draft Street Address Standard 2005, BS 7666:2006). As an example of a national address data grid, in South Africa, there would be roughly 260 address dataset publications, one for each municipality, and twelve addressing systems. Even if the above international and national figures are doubled or tripled, the resulting total numbers are still relatively small in respect of what relational DBMS, object-oriented DBMS and XML databases are able to cope with. The number of node hosts and registered address-related services are even more difficult to predict. All in all, however, one should be able to accommodate this information in a relational DBMS, an object-oriented DBMS or an XML database. There is no need to make special provision for huge volumes of data.

The structure of the data model for the address data catalogue is sufficiently simple to allow representation in a relational data model. There are no complex or semi-structured data modeling requirements in Compartimos that call for an XML or object-oriented data model. However, these two models are also possible technology choice. Since the catalogue is replicated among the nodes, it is important that the storage mechanism for the catalogue is platform independent so that it can be easily replicated at any node host. In light of this requirement, an XML data store is attractive.

The Compartimos catalogue includes two types of metadata:

1. Metadata about the address data, i.e. metadata about the address dataset publications, addressing systems and address-related services.
2. Metadata about the grid configuration, i.e. metadata about the replicas and the node hosts.

Using existing metadata standards holds the advantage that existing metadata can be readily imported into the Compartimos catalogue, and tools for capturing metadata according to these standards are also available. Metadata about the address datasets can be stored according to existing metadata standards, such as ISO 19115:2003*, Geographic information – Metadata*, with or without the ISO 19139:2007, *Geographic information - Metadata - XML schema implementation*. ISO 19119: 2005, *Geographic information – Services*, includes a data model for service metadata, which is applicable to the address-related services in Compartimos, while ISO 19112:2003 could be used

for the addressing systems, as described in Chapter 4. Metadata, and associated standards, is an important ingredient for an SDI and the above ISO standards are widely used in the geospatial community and in SDIs around the world (Aalders 2005). An alternative technology choice is the Dublin Core Metadata element set (www.dublincore.org), which has been adopted as an ISO standard (ISO15836:2003), but Dublin Core does not cater for spatial data specifically, and is not widely used in the geospatial community. The Globus Toolkit includes a catalogue capability (Singh *et al.* 2003, Zhao *et al.* 2004) but it does not address all the requirements for spatial data, such as, for example, the geographic extent of a dataset. However, the metadata that forms part of the Globus Toolkit's Replication Location Service (RLS) (http://www.globus.org/toolkit/data/rls/) would be an option for replica information in the Compartimos catalogue. The metadata that is part of the Globus Toolkit's Monitoring and Discovery System (MDS) (http://www.globus.org/ toolkit/mds/) provides information about the available resources on the Grid and their status and would be suitable to store information about the hosts in Compartimos. RLS and MDS were successfully integrated into a geospatial grid, as reported by Di *et al.* (2008).

5.2.2 The catalogue service (CatalogueService)

The operations of the Compartimos CatalogueService are listed in Table 40 of Appendix C, and give an idea of the capabilities that are required by this service. Due to the two types of metadata in the Compartimos catalogue, discussed in 5.2.1 above, the Compartimos Catalogue can be implemented as two separate services, each responsible for a particular type of metadata. The OGC has published a catalogue service implementation specification (OGC 2007) for the discovery and retrieval of metadata about spatial data and services. The OGC catalogue service can be implemented in conjunction with the above-mentioned ISO 19115 and companion standard ISO 19139, as well as ISO 19119 for service metadata.

Thus, the OGC catalogue service implementation specification is a technology option for the address-related catalogue service in an address data grid. Wei *et al.* (2006) and Di *et al.* (2008) report on using the OGC catalogue service in their implementation of a geospatial grid for NASA. Zhao *et al.* (2004) report on a different option in (seemingly) the same implementation of a geospatial grid for NASA, i.e. augmenting the Globus Toolkit's Metadata Catalogue Service (MCS) with the profile of the OGC Web Catalogue Service. Thus one can see that there is a need for the geospatial and grid communities to collaborate, so that the respective standards and tools can be used in together and duplication and overlap is minimized.

From a grid point of view, there are two relevant services in the Globus Toolkit, the Monitoring and Discovery System (MDS) (for hosts in Compartimos) and the Replica Location Service (RLS) (for dataset replicas in Compartimos).

5.2.3 The replica service (ReplicaService)

The operations of the ReplicaService in Compartimos are deliberately similar to the operations specified for a replica service in the OGSA data architecture, so that tools developed out of the OGSA community can be 'plugged' straight into Compartimos, because this is a generic service that does not require specialization for geographic data. The primary functionality of the Globus Toolkit's Data Replication Service (DRS) (http://www.globus.org/toolkit/docs/4.0/techpreview/datarep/) allows a user to identify a set of desired files existing in their Grid environment, to make local replicas of those data files by transferring files from one or more source locations, and to register the new replicas through the RLS.

An alternative is to integrate replication capabilities provided by whatever DBMS or GIS software runs at the data hosts. However, there might be syntactic and semantic data interoperability issues when replicating the 'raw' datasets from one DBMS or GIS database to another. Also, if proprietary DBMS or GIS software is used, the openness of the address data grid is compromised and licenses for the vendors' software have to be acquired at the relevant hosts where the data is replicated.

5.2.4 The transfer service (TransferService)

In Compartimos the purpose of the TransferService is to generically transfer data in the address data grid. The TransferService does not need to understand the data that is being transferred. Therefore this service is best suited to be 'plugged' in from other sources. The Globus Toolkit's Reliable File Transfer (RFT) service (http://www.globus.org/toolkit/docs/4.0/data/rft/) is a Web Services Resource Framework (WSRF) compliant web service that provides "job scheduler"-like functionality for data movement. One has to provide a list of source and destination URLs and then the service writes the job description into a database and then moves the files at a later stage. RFT makes use of GridFTP, a protocol defined by the Open Grid Forum and currently a draft before the IETF FTP working group. The GridFTP protocol provides for secure, robust, fast and efficient transfer of (especially bulk) data. The Globus Toolkit provides the most commonly used implementation of that protocol, though others do exist (primarily tied to proprietary internal systems). Di *et al.* (2008) and Wei *et al.* (2006) report on using the OGC catalogue service in their implementation of a geospatial grid for NASA.

5.2.5 The address data access service (AddressDataAccessService)

The OGC Web Feature Service (WFS), which returns spatial data in vendor independent GML format (ISO 19136:2007) is a natural choice for this service. This implementation specification is currently in the process of becoming adopted as an ISO standard, ISO 19142 (draft), *Geographic information – Web feature service*. However, additional functionality is required for the conversion

to and from the interoperable address data model specified by Compartimos. Aloisio *et al.* (2005a), Di *et al.* (2008), Wei *et al.* (2006) and Zhao *et al.* (2004) report on grid-enabling OGC web services, such as WFS and WMS.

An alternative technology choice is the OGSA-DAI software, which is compatible with the Globus Toolkit. However, OGSA-DAI has been developed for alphanumeric data and would require some extensions to accommodate spatial data. However, OGSA-DAI resources are already usable by other Globus Toolkit services. The choice of OGSA-DAI would also influence the technology choice for other services such as the CatalogueService and the VirtualAddressDataService.

The third option is to integrate OGC web services with OGSA-DAI as was reported by Shu *et al.* (2004). This option would have the same benefit of seamless integration with the Globus Toolkit as the alternative above.

5.2.6 The virtual address data service (VirtualAddressDataService)

The VirtualAddressDataService is the center of intelligence in Compartimos. This is where distributed queries are executed, where resulting data is consolidated, where duplicate addresses are removed, and where address disambiguities are resolved, to name a few. Since these are diverse capabilities each within its own field of specialization, it will make sense to combine different components for the implementation of the VirtualAddressDataService. The potential combinations are huge, but the following are examples:

1. the OGSA-DAI Distributed Query Processing (DQP) (http://www.ogsadai.org/about/ogsa-dqp/) could be employed for distributed queries (with the implication that this influences the technology choice of other services);
2. the address matching functionality provided by independent tools such as the AfriGIS Intiendo address tool (Rahed *et al.* 2008) could be used to remove duplicates and resolve disambiguities; and
3. any processing that is available as a OGC Web Processing Service (WPS), which is a standardized interface that facilitates the publishing of geospatial processes, and the discovery of and binding to those processes by clients. Since WPS is on the initial list of goals for grid integration included in the OGC/OGF MoU, this technology choice is relevant.

Di *et al.* (2008) implemented their own mediator for geographic data, the Intelligent Grid Service Mediator (iGSM), while Shu *et al.* (2004) propose using OGSA-DAI DQP. Once the address data has been consolidated, similar to the AddressDataAccessService, an implementation of WFS or OGSA-DAI data resources are potential technology choices.

5.2.7 The address dataset (AddressDataset)

In Compartimos the data provider determines how address data is stored. The AddressDataAccessService provides access to this proprietary data in the prescribed way (according to the interoperable data model). However, for optimal conversion efficiency it will make sense to store the 'raw' data according to the Compartimos interoperable data model, or as close to it as possible.

5.2.8 The address-related service (AddressService)

The functionality and interface of this service is determined by its purpose, and therefore not prescribed in Compartimos. For interoperability, it is important that this service uses the same standard and protocol as the other services in Compartimos. The OGC WPS would be a standardized choice for integrating third party address-related services, such as geocoding or routing, into Compartimos. Alternatively, depending on the purpose of the service, the OGC WFS could also be used.

5.3 Overall technology choices for Compartimos

In this section overall technology choices, relevant to Compartimos as a whole, are discussed. These choices have an impact on the technology choices for the individual Compartimos objects, and should thus not be evaluated in isolation.

5.3.1 Security

The most obvious technology choice for an address data grid would be the Globus Toolkit's Grid Security Infrastructure (http://www.globus.org/toolkit/docs/latest-stable/security/). GSI is concerned with establishing the identity of users and/or services (authentication), protecting the integrity and privacy of communications (message protection), determining and enforcing who is allowed to perform what actions on what resources (authorization), and provide (secure) logs to verify that the correct policy is enforced (accounting allows for auditing of policy compliance). GSI is based on the standard X.509 end-entity and proxy certificates, which are used to identify persistent entities such as users and servers and to support the temporary delegation of privileges to other entities. Di *et al.* (2003), Aloisio *et al.* (2005a, 2005b) and *Wei et al.* (2006) use the GSI in their respective implementations of geospatial grids.

5.3.2 Operating system and/or programming language

Any implementation of Compartimos has to accommodate a distributed heterogeneous environment and therefore, in principle, any operating system is acceptable for the hosting environment for each of the data, service and node hosts. Flavours of UNIX and Windows are

probably the most obvious choices. This implies that the services, as well as the catalogue, have to be portable onto the different operating systems so that they can be deployed on any operating system. In practice, however, there might be restrictions on the number of platforms that are supported at the various hosts. Table 14 lists the Compartimos services and the respective hosts where they are deployed.

The choice of platform for the individual AddressDataset and associated AddressDataAccessService on the *data host* lies solely with the respective owners. The AddressDataset can be implemented on any platform, since the AddressDataAccessService provides the platform independent access to the AddressDataset. The choice of platform and/or programming language for the AddressDataAccessService is also open, as long as it provides a platform independent communication interface as a web service. The same holds for the AddressService on the *service host* that can be implemented on any platform in any programming langage, as long as it provides a platform independent communication interface as a web service.

Table 14. Where Compartimos services are hosted

Compartimos object	Description	Hosted on
AddressDataAccessService	Provides uniform access to individual address datasets	Data host
AddressService	A third party address-related service such as routing or mapping	Service host
AddressDataset	The individual address dataset	Data host
CatalogueService	Provides read and update access to the catalogue	Node host
ReplicaService	Replicates data in the address data grid	Node host
TransferService	Transfers large volumes of address data	Node host
VirtualAddressDataService	Consolidates the data	Node host

With the services that run on the *node host*, i.e. the CatalogueService, ReplicaService, TransferService and VirtualAddressDataService, the choice of operating system and programming language has a bigger impact. In principle, each one of these services can be implemented in a different programming language in a different operating system. The node would then host these services in different virtual operating systems, and they woud communicate with each other as platform independent web services. In reality, it might be simpler to restrict a node host to a single operating system on which the above-mentioned Compartimos services are deployed. If the services are implemented in a platform independent programming language such as Java, they can be easily ported to run on different operating systems, thus enabling individual node hosts to run on different operating systems.

The OGC web service implementation specifications, proposed as technology choices for Compartimos objects in the previous section, are able to communicate with each other, regardless of the platform.

5.3.3 Web service protocols

Platform independent protocols ensure that the Compartimos services are able to communicate even though they are deployed on hosts with different operating systems. The W3C and OASIS standards specify communication through XML messages that follow the SOAP protocol. Another option is RESTful (representation state transfer) web services, referring to a simple interface, which transmits domain-specific data over HTTP without an additional messaging layer such as SOAP. It is beyond the scope of this dissertation to provide a full comparison of pros and cons of these two kinds of web services, but it is recommended that a single kind of web service be used within a specific Compartimos implementation.

The choice of communication protocol is also related to available standards. The OGF, where grid standardization takes place, are cooperating with OASIS and thus web services based on SOAP are prevalent. The OGC Web Service Implementation Specifications on the other hand, is based on POST/GET methods through HTTP. This poses problems for any integration of OGF and OGC web services. However, a recently completed OGC Web Services, Phase 5 (OWS-5) Testbed, included 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 this matter.

5.3.4 Connectivity and bandwidth

All Compartimos hosts should be able to connect to the Internet. However, it is possible that some data hosts are not continuously connected. Some data hosts might have 24-hour connectivity, while others might connect to the Internet with the specific purpose of uploading and synchronizing a dataset. Such a data host would initially upload the dataset, which is then replicated at a node, and from then on synchronized by the data host at regular intervals. Theoretically, it is necessary that at least one node in the grid has 24-hour connectivity so that the catalogue and at least one VirtualAddressDataService are always available. In practice, the number of data requests will determine the number of nodes and associated configurations (i.e. which services they host) that are required.

Nodes hosts should have sufficient bandwidth to be able to handle the data transfers required in Compartimos. The amount of bandwidth required depends on the size of individual datasets and granularity of the replication strategy.

5.4 Proof of concept implementation of Compartimos

Compartimos has been implemented as a proof of concept in a controlled environment on a single computer at the University of Pretoria. In this controlled environment hosts (data, service and nodes) are simulated as individual web applications. In this implementation all communication is

through the web server and standard Internet protocols, so that the implementation is easily transferable to a number of web applications distributed across a number of geographically distributed machines in different administrative domains, connected to the Internet and collectively comprising an address data grid. These distributed web applications would typically reside on the servers of individual SDI participants, such as individual local authorities.

The Compartimos address data grid is accessible through web services to any external application, and in the controlled environment this external programmatic access is illustrated with a graphical user interface (GUI) for a website portal. The portal gives access to the catalogue of the address data grid, and also implements the three use case scenarios that were described in Chapter 4 for a simple data request, an iterative data request and an address-related service by a third party. The three use cases thus interact with the CatalogueService, the VirtualAddressDataService and an AddressService. Figures 40-42 show some screenshots from this portal: the home page, address data results and catalogue data. 'NAD on the Grid' in the logo refers to the THRIP project that is jointly funded by the South African Department of Trade and Industry and AfriGIS, a South African GIS service provider, and of which this research is a part (refer to the Preamble and Acknowledgements).

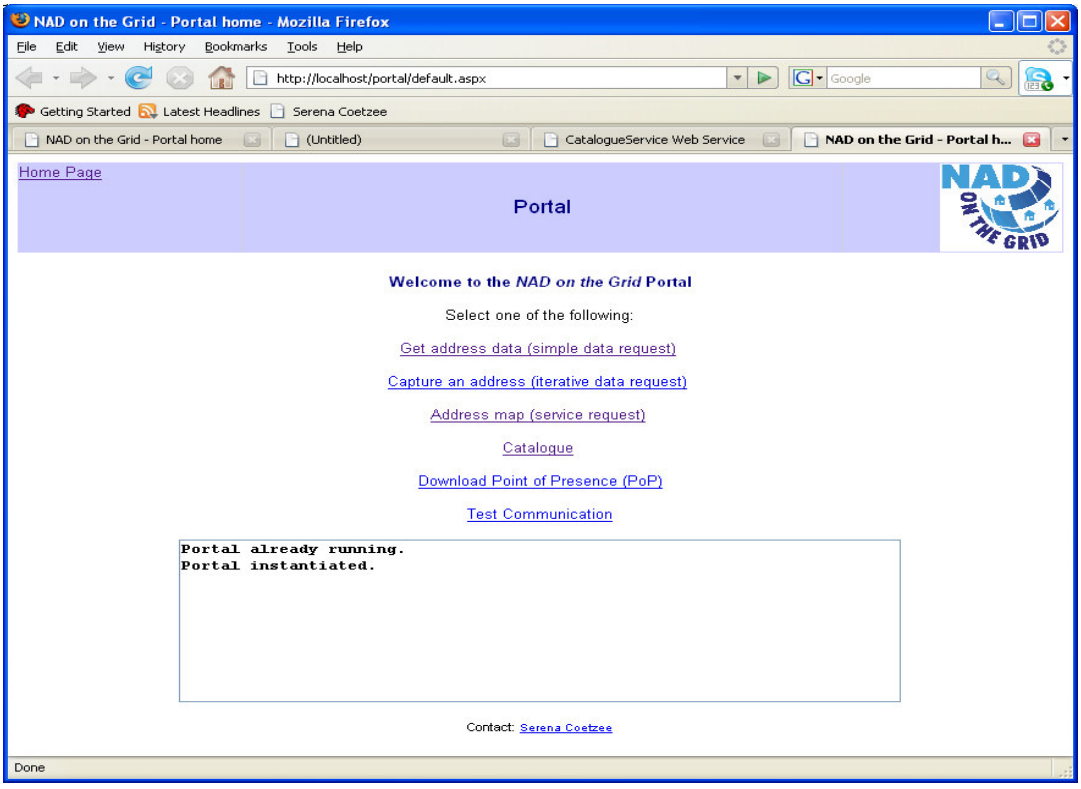
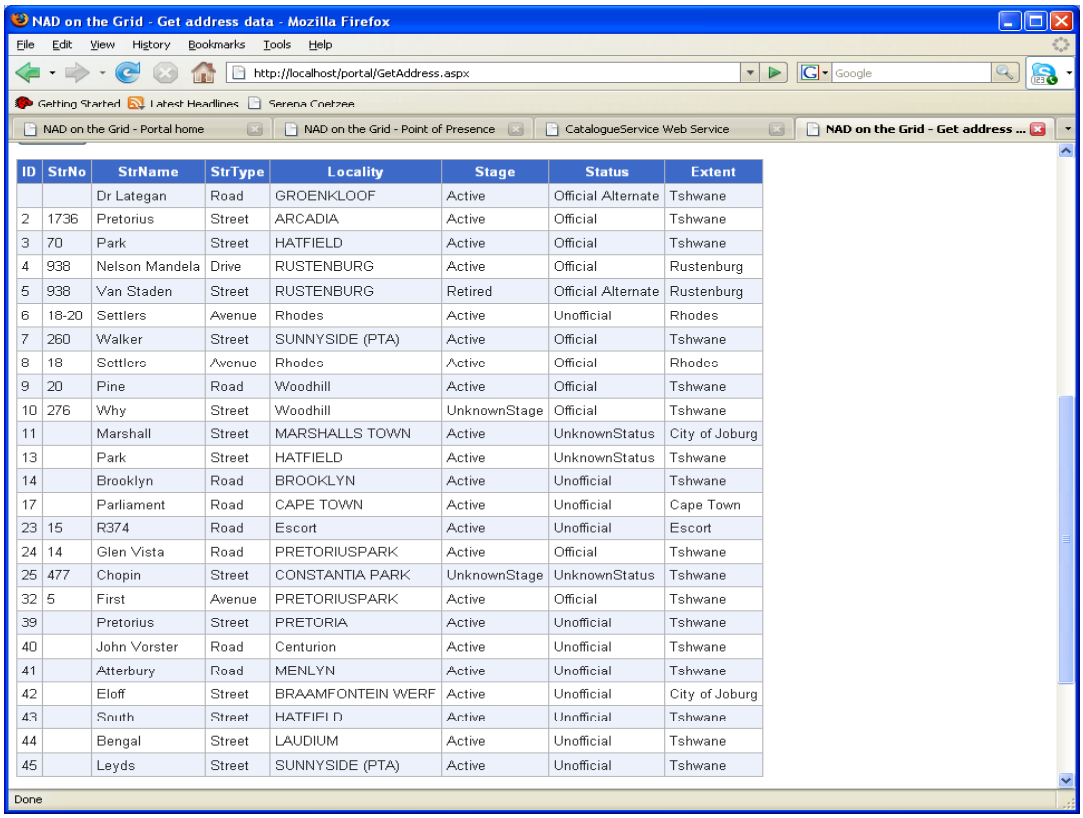


Figure 40. Home page of the address data grid portal

The portal and web services are implemented in C# on the Microsoft Windows platform. The

catalogue data was stored in a Microsoft Access database, which can quite adequately process the current small volumes of catalogue information in the controlled environment. The interface to the catalogue is isolated in a separate class, which can easily be replaced with a class that accesses data in a different relational DBMS, or other platform independent data repository. Therefore, at this early stage, the platform dependence of Microsoft Access is not considered problematic in the controlled environment.



ID	StrNo	StrName	StrType	Locality	Stage	Status	Extent
		Dr Lategan	Road	GROENKLOOF	Active	Official Alternate	Tshwane
2	1736	Pretorius	Street	ARCADIA	Active	Official	Tshwane
3	70	Park	Street	HATFIELD	Active	Official	Tshwane
4	938	Nelson Mandela	Drive	RUSTENBURG	Active	Official	Rustenburg
5	938	Van Staden	Street	RUSTENBURG	Retired	Official Alternate	Rustenburg
6	18-20	Settlers	Avenue	Rhodes	Active	Unofficial	Rhodes
7	260	Walker	Street	SUNNYSIDE (PTA)	Active	Official	Tshwane
8	18	Settlers	Avenue	Rhodes	Active	Official	Rhodes
9	20	Pine	Road	Woodhill	Active	Official	Tshwane
10	276	Why	Street	Woodhill	UnknownStage	Official	Tshwane
11		Marshall	Street	MARSHALLS TOWN	Active	UnknownStatus	City of Joburg
13		Park	Street	HATFIELD	Active	UnknownStatus	Tshwane
14		Brooklyn	Road	BROOKLYN	Active	Unofficial	Tshwane
17		Parliament	Road	CAPE TOWN	Active	Unofficial	Cape Town
23	15	R374	Road	Escort	Active	Unofficial	Escort
24	14	Glen Vista	Road	PRETORIUSPARK	Active	Official	Tshwane
25	477	Chopin	Street	CONSTANTIA PARK	UnknownStage	UnknownStatus	Tshwane
32	5	First	Avenue	PRETORIUSPARK	Active	Official	Tshwane
39		Pretorius	Street	PRETORIA	Active	Unofficial	Tshwane
40		John Vorster	Road	Centurion	Active	Unofficial	Tshwane
41		Atterbury	Road	MENLYN	Active	Unofficial	Tshwane
42		Eloff	Street	BRAAMFONTEIN WERF	Active	Unofficial	City of Joburg
43		South	Street	HATFIELD	Active	Unofficial	Tshwane
44		Bengal	Street	LAUDIUM	Active	Unofficial	Tshwane
45		Leyds	Street	SUNNYSIDE (PTA)	Active	Unofficial	Tshwane

Figure 41. Results of a simple data request displayed in the portal

AfriGIS address data was used for the proof of concept implementation. A number of smaller datasets were extracted from the national address dataset, each representing the address data that would typically come from an individual local authority in an SDI. The individual datasets include addresses with various address types from a number of areas in South Africa. The focus was on testing the capability of address dataset coordination and consolidation, rather than on transferring large volumes of address data.

All the Compartimos objects are implemented in the proof of concept: the Catalogue, CatalogueService, AddressDataset, AddressDataAccessService, VirtualAddressDataService,

ReplicaService, TransferService and AddressService. However, the main goal was to get the first five objects, those with address-related capabilities, functional. The replication and data transfer functionality of the *ReplicaService* and *TransferService* is generic (specialization for address data is not required) and the Reliable File Transfer (RFT) service and the Data Replication Service (DRS) from the Globus Toolkit, for the *ReplicaService* and *TransferService* respectively, will be considered when expanding the implementation.

The purpose of the implementation was to investigate and experiment with the architectural aspects of Compartimos, i.e. the list of constituent objects and their capabilities. The controlled environment served this purpose well: it was quick to test different combinations of services with different capabilities without having to go through the administrative and human communication of orchestrating hosts at different sites. The result of these experiments is the Compartimos reference model, presented in Chapter 4.

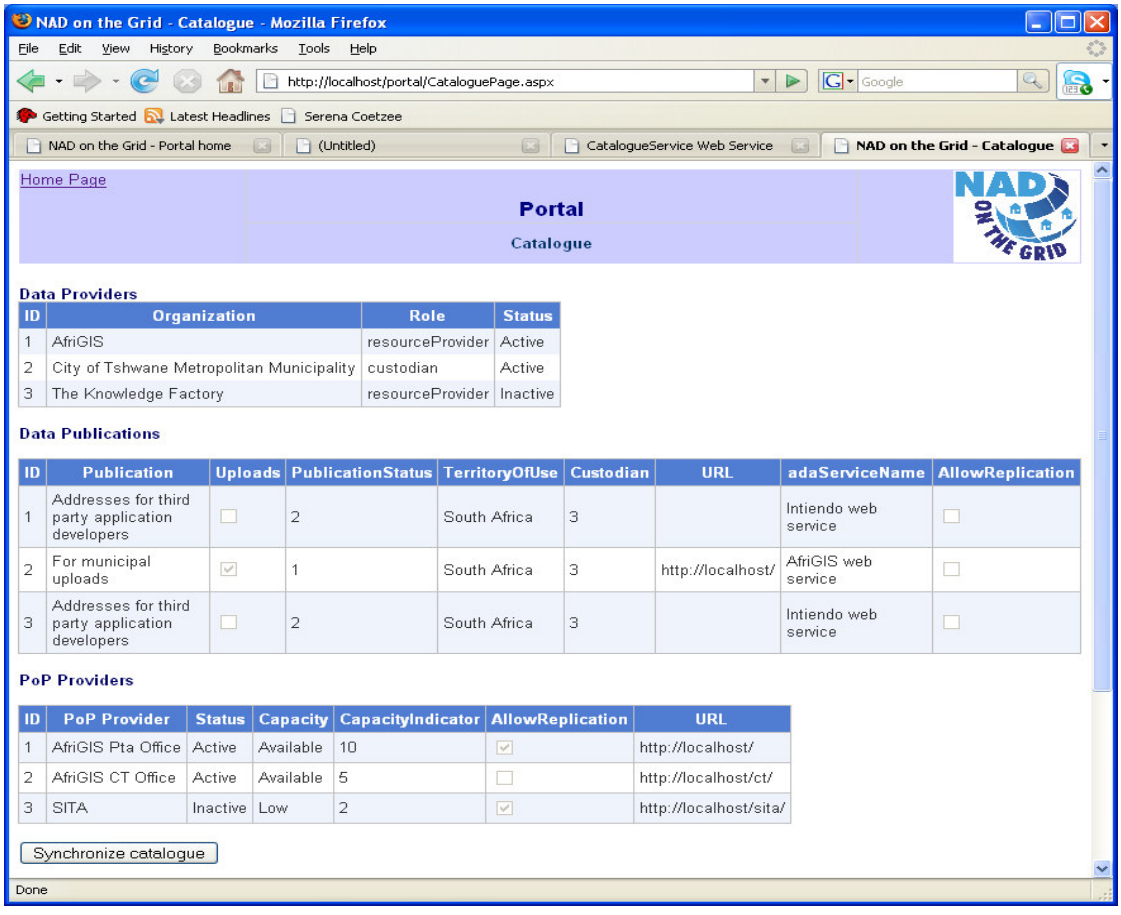


Figure 42. Catalogue contents displayed in the portal

Now that Compartimos has been finalized, the next steps are to experiment with larger datasets and to physically distribute the servers. This next round of investigations would focus on the efficiency of Compartimos, rather than on its constituent objects and their list of capabilities. In a potential second phase additional data, service and node hosts can be deployed on servers at the university and at AfriGIS, one of the sponsors of this research project, in order to test Compartimos' efficiency with large address datasets. While this will still be a closed and controlled environment, such a roll out will also have to make provision for security in Compartimos, and GSI will be considered for this. As a potential third phase hosts can be deployed at one or two local authorities, at the State IT Agency (SITA) and other SDI stakeholders, thus making the address data grid available in an SDI environment.

5.5 Evaluation of Compartimos

This section provides a retrospective look on the Compartimos reference model. Firstly, Compartimos is compared against the novel evaluation framework for national address database in an SDI that is presented in Chapter 6. Secondly, results from the development of Compartimos are discussed.

5.5.1 Evaluation against criteria of the evaluation framework

In this section the Compartimos reference model is evaluated against the novel evaluation framework for national address databases that is presented in Chapter 6. Even though the evaluation framework is presented at the end of this dissertation, it was developed during the initial stages of research to establish whether a data grid approach would be suitable for national address databases in an SDI. The framework comprises of criteria that are based on the requirements for the establishment, maintenance and use of a national address database. As part of the discussion of Compartimos in this chapter, it is appropriate to evaluate the Compartimos reference model against the criteria in this framework.

Tables 15-21 below show the criteria from the novel evaluation framework and refer to the relevant section in this dissertation where the criterion is addressed. There are two criteria that are not met: *Feedback from users to data providers* (in Table 18) and *Billing and accounting* (in Table 19). It is suggested that feedback is addressed as part of further research, refer to 7.3.3. Methods for charging for address data in a national address database were researched separately and still have to be integrated into Compartimos (Acton 2007). All in all, the evaluation in the tables below confirms that Compartimos will address the requirements for national address databases in an SDI.

Table 15. Infrastructure

Criteria	Description	Compartimos
Operating system	Data and service providers should be free to use the operating system of choice	Refer to the technology choices for the operating system in 5.3.2
Database Management System (DBMS)	A data provider should be free to store the address data in a DBMS (Oracle, SQL Server, ArcSDE, ESRI SHP files, MapInfo files, etc.) of choice	Refer to the technology choices for the catalogue in 5.2.1, the address dataset 5.2.7 and the operating system in 5.3.2
Address format	Although address-related services should be based on a standardized address format, the unified view layer should accommodate the differences in address representation of the individual data providers	Refer to the address data model in Chapter 4, as well as the address data access service in 5.2.5

Table 16. Data providers

Criteria	Description	Compartimos
Coverage area	Variation in the size and location of the coverage of address databases supplied by data providers should be allowed, and data access should be optimized for this, i.e. don't search for a Cape Town address in the Johannesburg database.	Refer to the technology choices for the virtual address data service in 5.2.6. Optimization will depend on the specific technology choice
Decentralized source of data	The reality of many decentralized sources of address data providers must be catered for.	Refer to Chapter 3 for a definition of the data grid, also the catalogue in 5.2.1
Multiple data providers per area	A data request should consider addresses from all the data providers, and resolve duplicates, ambiguities and potential semantic differences.	Refer to the technology choices for the virtual address data service in 5.2.6

Table 17. Naming

Criteria	Description	Compartimos
Suburb Names	Enough information (such as alias tables) as well as disambiguating functionality should be provided to resolve between new official and colloquial names for suburbs.	Refer to the technology choices for the virtual address data service in 5.2.6. Efficiency of this capability will depend on the specific technology choice.
Name Changes	Enough information (such as alias tables) as well as disambiguating functionality should be provided to resolve between new and old names of suburbs and streets.	

Table 18. Address Dynamics

Criteria	Description	Compartimos
New developments	Address data for newly developed areas should become available as soon as possible. A quarterly update cycle is too long.	Because address data is stored at data hosts close to a local authority, the latest data is available in the data grid. Refer to deployment options in the engineering viewpoint (4.5)
Previously un-addressed	Newly assigned addresses in previously unaddressed areas should be accessible as soon as possible in order to speed up service delivery to the areas as part of the development initiative in a country.	
Address cross checking	Data providers should be able to cross check the availability of address data in areas for which they plan to produce address data.	Data providers can make use of the virtual address data service (5.2.6) to check availability of data.
Feedback from users to data providers	Users of the address data should be able to provide feedback to data providers about the correctness and accuracy of address data.	The Compartimos reference model does not include this capability, but it is one that can easily be added.

Table 19. Accessibility

Criteria	Description	Compartimos
Providing services (service providers)	Service providers should be able to provide value-adding address-related services on top of the unified view of the national address data. These services should be provided in a standard and well-known framework such as web services, and more specifically web feature services as specified by the Open Geospatial Consortium (OGC).	Refer to the address-related service in 4.4.9 and 5.2.8
Billing and Accounting	The information federation model should allow a two-level billing and accounting system for both data use, and the use of vendor-supplied services.	Compartimos does not yet provide for billing, although this was researched in a separate project (Acton 2007)
Using services (application developers)	Application developers should be able to seamlessly integrate into their applications both services that provide access to the unified view of the national address database as well as the vendor-supplied services.	Refer to the address-related service in 4.4.9 and 5.2.8, as well as the virtual address data service in 5.2.6
Access anytime	Access through these services to the national address database should be instantaneous and available all the time.	Refer to the technology choices for connectivity and bandwidth in 5.3.4
Access from anywhere	Access to the national address database should be available from as many platforms as possible including client desktops, personal digital assistants (PDA) and/or mobile phones.	Refer to the technology choices for operating systems in 5.3.2
Ease of publishing data (providing data)	Facilities for publishing address data should be easy and should not require specialized IT support.	Refer to the technology choices for the catalogue service in 5.2.2, and also the GUI implementation of the catalogue in 5.4

Table 20. Security

Criteria	Description	Compartimos
User Authentication	Access to the national address database should be restricted to authenticated users.	Refer to the technology choices for security in 5.3.1
Access	Data providers should be able to specify how and to whom (which group of people) their data is available.	
Privacy	The data in the national address database should be protected against unauthorized access.	

Table 21. Organizational Issues

Criteria	Description	Compartimos
Official custodians and unofficial data providers	The information federation model for a national address database should support the fact that there could be both officially regulated address data providers, supporting an official national address register, and unofficial address data providers, supporting national address databases in general.	Refer to the technology choices for the virtual address data service in 5.2.6, as well as the catalogue in 5.2.1

5.5.2 Discussion of results

In this section results and conclusions forthcoming from the work on Compartimos are discussed. Recommendations for future research are presented in Chapter 7.

The OGSA data architecture describes the interfaces, behaviors and bindings for manipulating data within the broader OGSA architecture, and Compartimos is based on this architecture. This implies that Compartimos follows the same service-oriented approach adopted in both OGSA and the OGSA data architecture. Where applicable, Compartimos provides the details of these services to make provision for address data in an SDI environment. Compartimos thus is an application domain-specific application of the OGSA data architecture, which could also be referred to as a 'profile' or specialization of the OGSA data architecture for address data in an SDI. In this dissertation the essential components required for an address data grid in an SDI environment are presented in the Compartimos reference model, and the required capabilities such as address interpretation and address consolidation are designated to specific Compartimos services. This designation distinguishes those parts of Compartimos that are application domain-specific from those that are not, thus identifying the Compartimos objects that have to be implemented for different application domains (in contrast to the generic ones).

The interoperable address data model that is proposed as part of Compartimos in Chapter 4 is one way of enabling address data interoperability. With this model, as soon as an addressing system is registered in the catalogue, address data that is based on that addressing system is 'understood' in the data grid. This model allows any number of addressing systems to be integrated into the data grid

so that a global model is not enforced on local data providers. However, the flipside of the coin is that there could be so many addressing systems, each ever so slightly different, that interoperability is not really achieved. The model works well if most data providers share a reasonable number of addressing systems. Hong (2008) proposes alternative ways of dealing with different kinds of location (address) representations, such as a logical representation for a location, a framework for conversion between different reference systems, and Compartimos might benefit from some of these ideas. Making use of the context, i.e. the characteristics of the user's environment, to interpret an address is another possibility. Brovelli *et al.* (2008) described how context-awareness could provide a richer user experience by adapting the user interface on a mobile device in relation to the context; similarly and address could be interpreted in relation to its context, for example, after entering only a street name and number on a mobile device, the suburb and higher level address information is derived from the position of the mobile device.

The purpose of the proof of concept implementation was to investigate the architectural aspects of Compartimos and the controlled environment served this purpose well. There are however aspects of Compartimos that cannot be tested in the controlled environment and require further investigation. For example, small datasets of addresses were used and the focus was on address dataset coordination and interpretation, rather than on transferring large volumes of address data. A data transfer service, such as Globus Toolkit's RFT that makes use of the GridFTP protocol and already widely used in the grid community for data volumes far larger than those required for address data, should be able to fulfill this requirement of Compartimos. Security was also not fully investigated in the controlled environment and should get more attention.

The technology choices for the address data access service show that there is more than one approach using OGSA-DAI and/or OGC web services. The reports on current grid-enablement of OGC web services provide further evidence that there is a need for grid-enabling OGC web services, and it shows that different approaches are possible. Standardization, if necessary, of technologies for these approaches should take place in the joint OGC-OGF forum. Further the technology choices regarding grid components are mainly from the Globus Toolkit since this is the de facto standard, open source and thus feely available. There are however other tools such as Alchemi, an open source product from the University of Melbourne (Arefin *et al.* 2006) and also products from Oracle, IBM, Sun and HP (Buyya and Nadiminti 2006).

In Compartimos the node hosts have been designed to be configurable in terms of the combination of components that they host, ranging from data hosts that upload data to the grid at intervals, data hosts that continuously provide access to data, to 'power' nodes that host all the components of the reference model. While it is expected that there is a requirement for these kinds of hosts, as justified in the description of the enterprise viewpoint, further confirmation will be

forthcoming when hosts are deployed at selected local and national authorities in potential subsequent phases. Another possibility altogether is to let these hosts live in a 'cloud', such as the Amazon EC2.

This research has focused on the technical aspects of an SDI, as discussed earlier, and therefore Compartimos is a technical reference model. How such a model should be implemented in an SDI from a viewpoint of human resources, legislation, policies, etc. is beyond the scope of this research, and requires further investigation.

Chapter 6 Address databases for national SDI: Comparing the novel data grid approach to data harvesting and federated databases

This chapter was published online (ahead of the print edition) on 26 September 2008 in the International Journal of Geographic Information Science (IJGIS), as a paper by Coetzee S and Bishop J under the same title.

6.1 Introduction

The original purpose of addresses was to enable the correct and unambiguous delivery of postal mail. The advent of computers and more specifically geographic information systems (GIS) opened up a whole new range of possibilities for the use of addresses, such as routing and vehicle navigation, spatial demographic analysis, geo-marketing, and service placement and delivery. Such functionality requires a database which can store and access spatial data effectively. In this chapter we present address databases and justify the need for national address databases. We describe models used for national address databases, and present our evaluation framework for an address database at a national level within the context of a spatial data infrastructure (SDI). The models of data harvesting, federated databases and data grids are analyzed and evaluated according to our novel framework, and we show that the data grid model has some unique features that make it attractive for a national address database in an environment where centralized control and/or coordination is difficult or undesirable.

A hundred years ago addresses were used mostly for postal delivery and land administration: national postal services used them for letter and parcel delivery and the deeds registry needed them to correctly and unambiguously record property ownership. The advent of computers, and more specifically geographic information systems (GIS), opened up a whole new range of possibilities for the use of addresses, such as routing and vehicle navigation, spatial demographic analysis, geo-marketing, service placement and delivery, and electronic address verification, to name a few. The

efficient and effective use of addresses in this way relies on the presence of a database that handles addresses in a spatial, rather than just a textual context.

In many countries address data producers operate on a local (town, county, local authority) level and their data has to be combined in various ways in order to provide access to an integrated national address database. In South Africa, for example, to gain access to integrated national address data one has to buy the dataset or subsets thereof from a limited number of private vendors. The cost of this data does not always justify buying it, and therefore one of the goals of our research to date is to investigate ways of providing address-related services rather than the address data itself. Our research also explores ways of providing integrated access to the various distributed address datasets thereby enabling independent service providers to provide address-related services with national and even international coverage. Integrating information from a number of heterogeneous databases into a single conceptual database is commonly referred to as information federation (Sheth and Larson 1990). We have developed a novel evaluation framework, which we use to evaluate three information federation models that could be applicable. Although our evaluation is set in the South African context, the work has global relevance. The following three sections discuss spatial data infrastructures (SDIs), national address databases (NAD) and data grids and how we combine them in this chapter. We conclude the introduction with an outline of our chapter.

6.1.1 Spatial Data Infrastructure (SDI)

Spatial Data Infrastructure (SDI) refers to the technologies, standards, arrangements and policies that are required to collate spatial data from various local databases, and to make these collated databases accessible and usable to as wide an audience as possible (Jacoby *et al.* 2002). National spatial data infrastructures emerged in the early 1980s in countries such as the USA and Australia. These first generation SDIs mostly followed a product-based approach. The next generation of SDIs is moving towards a more process-based approach focusing on the creation of a suitable infrastructure to facilitate the management of information access, instead of the linkage to existing and future databases (Crompvoets *et al.* 2005). Web services are a prominent and important feature of these process-based SDIs. Masser *et al.* (2007) further point out that the concept of an SDI is evolving from being a mechanism of data sharing to becoming an 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. SDI as an enabling platform can be viewed as an infrastructure linking people to data on the basis of the common goal of data sharing.

6.1.2 National Address Database (NAD)

A national address database (NAD) falls into the realm of a country's spatial data infrastructure. In the preparatory work of the European program for an SDI, INSPIRE (INfrastructure for SPatial

InfoRmation in Europe), 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 requirements and have therefore been included explicitly in the final INSPIRE Directive in 'Annex 1', which contains the priority spatial reference datasets. This importance of address data as address data is applicable in other countries as well.

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 (Coetzee *et al.* 2008b). However, the need for address data for areas that extends across these jurisdictional boundaries calls for the collation of address data on a national and/or international scale. Example implementations of national address databases in Australia and Ireland follow the data harvesting model where all local data is loaded into a single centralized database that is under the control of a single organization.

6.1.3 Data grids

Grid computing started in the late 1990s as a distributed infrastructure for specific Grand Challenge applications executing on high-performance hardware. Since those initial days, it has evolved into a seamless and dynamic virtual environment (Baker *et al.* 2005). Although the initial focus of grid computing was on computational performance, it has expanded to address the needs of virtual organizations providing flexible, secure, coordinated resource sharing among collections of individuals, institutions and resources (Foster *et al.* 2001). There are different categories of grids such as computational grids, access grids and data grids, the last being the focus of this study. Data grids primarily deal with providing services and infrastructure for distributed data-intensive applications. Venugopal *et al.* (2006) identified a few unique features of 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 features are similar to the challenges facing the development of a national SDI as 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). They also correspond to the 'federation-by-accord' data sharing model mentioned by Harvey and Tulloch (2006). Thus there is a pre-existing link between the background to SDI and data grids, which we explore in this chapter.

6.1.4 Combining SDIs, NAD and data grids

The importance of address data as reference data together with the fact that address data is usually maintained on a local level but required on a larger scale implies that the principles of SDIs apply to the collation of address data into a NAD. 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 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. Harvey and Tulloch (2006) describe some disadvantages to giving a single organization the authority over data production and sharing and report that a federation-by-accord, although difficult to establish, once integrated into ongoing activities, can become sustainable and a suitable vehicle for enhancing data sharing. Our novel approach to a national address database as a data grid corresponds to the 'federation-by-accord' data sharing model which can afford to lose a major player without ruining the entire model.

In this chapter we explore three information federation models that could potentially support this 'federation-by-accord' data sharing model: data harvesting, federated databases and data grids. The large number of organizations involved in a national address database, as well as the lack of a single organization tasked with the management of a national address database, presents the data grid as an attractive alternative to the other two models. The data grid provides for a more loosely coupled architecture, thereby allowing for more diversity and heterogeneity. Both the data harvesting model and the federated database model require a single organization to take control. Our novel approach to a national address database as a data grid corresponds to the 'federation-by-accord' data sharing model, which can afford to lose a major player without ruining the entire model.

6.1.5 Outline of the chapter

The chapter is divided into four sections. In section two we present the status of address data and justify the need for address databases at a national level. Section 3 describes our novel evaluation framework that is used to evaluate three information federation models. In section four we discuss three models for federation of information: data harvesting, a federated database, and a data grid. We analyze the models by comparing their purpose, how the unified view of the integrated data is established, how data updates are done, and whether transactions and service-orientation are supported.

In section five we evaluate and analyze the three models according to our novel evaluation framework and describe some implementation issues. The analysis of the three models shows that where a large number of organizations are involved, such as for a national address database, and where there is a lack of a single organization tasked with the management of a national address

database, the data grid is an attractive alternative to the other two models. The data grid provides for a more loosely coupled architecture, thereby allowing for more diversity and heterogeneity. We explore this novel data grid approach to a national address database and also point out how this supports other decentralized approaches such as the 'federation-by-accord' data sharing model.

In summary, the objectives and contributions of this chapter are to 1) sketch the status of spatial address data within the context of a SDI in a country like South Africa; 2) present our novel evaluation framework for national address databases; 3) describe potential information federation models for national address databases; and 4) evaluate these models according to our evaluation framework.

6.2 Spatial address data

6.2.1 Address data

We define an address as a code or description for the fixed location of a home, building or other entity, and a spatial address as an address together with a coordinate for the geo-referenced location of the address. Our definition of an address does not include any information about the person or business residing at the address. Table 22 below lists sample addresses from a number of countries.

Table 22. Sample Addresses

Country	Address	Country	Address
Germany	Waldparkstrasse 67c DE-22605 Hamburg GERMANY	Spain	Calle Agazado, 23 Molino de la Hoz Las Rosas ES-28230 MADRID SPAIN
Japan	14F Sphere Tower Tennoze 2-2-8 Higashishinagawa Shinagawaku Tokyo 140 0002 Japan	Turkey	27 Gül Sokak 61250 Yomra Trabzon Turkey
New Zealand	6 Upland Road Kelburn Wellington 6005 New Zealand	United Kingdom	Russell House 4395 Station Road Porchester FAREHAM PO16 8BQ

A spatial reference system is used to identify locations on the surface of the Earth and addresses in an addressing system can be described as locations in a spatial reference system (Coetzee *et al.* (2008b). There are three types of reference systems:

1. a coordinate reference system specifies the location by reference to a datum;

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; and

3. a geographic identifier reference system specifies the location by a label or code.

According to ISO 19112, *Geographic information - Spatial referencing by geographic identifiers*, a geographic identifier reference system comprises a related set of one or more location types, that may be related to each other through aggregation or dis-aggregation, possibly forming a hierarchy. Davis and Fonseca (2007) conclude that this notion of an address as a hierarchy is commonly found in addressing systems around the world. An example of a geographic identifier reference system in South Africa would be Country > Province > Municipality > Suburb; and a location instance in this system would be South Africa > Gauteng > City of Tshwane Metropolitan Municipality > Hatfield. The similarity between a geographic identifier reference system and an addressing system can be illustrated by extending the geographic identifier reference system to include street names and street numbers, as in Country > Province > Municipality > Suburb > Street > Street Number. This allows a street address to be represented as a location instance of this reference system: South Africa > Gauteng > City of Tshwane Metropolitan Municipality > Hatfield > Pretorius Street > 1083. The British address standard, BS 7666, was developed in line with this notion of a geographic identifier reference system, proving that an addressing system can be viewed as a geographic identifier reference system.

However, if address numbers are assigned according to distance, then thoroughfare addressing can be regarded as a type of linear referencing system, as opposed to a geographic identifier reference system. For example, if address numbers are increased at one per meter, then '310 King Street' means: 'Proceed 310 meters along King Street from its beginning, then look to the even-numbered side of the street', i.e. route, reference point, distance, offset.

In the extreme case, addressing can even resemble a coordinate reference system. For example, in South Africa addresses in remote rural areas are captured as 'dots' either with GPS devices or from aerial photography. Each one of these dots represents an address. The geographic identifiers associated with the dot could include the province, municipality and village name, but no more than that. To locate the address, one has to know the coordinate. Over time these addresses could evolve into addresses, as we more commonly know them, i.e. including street names and numbers.

Thus, one can consider addressing to fall into all three types of reference systems, or consider addressing to be a fourth type of reference system due to the potential many-to-many relationships between, for example, an address and what is being addressed such as a building or a land parcel.

The importance of address data as reference data is illustrated in the preparatory work of the European program for an SDI, INSPIRE (INfrastructure for SPatial InfoRmation in Europe), 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 requirements: in numerous legacy and modern IT systems, address information is recorded with the purpose of having an unambiguous identification of the real property, 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 property or insurance. Last but not least, 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 the final INSPIRE Directive in 'Annex 1', which contains the priority spatial reference datasets.

The typical responsibilities of local governments often cause them to 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. There is often little or no cooperation between local and national government, and the trend to manage and maintain the national address database by adding local data to a single centralized database and periodically publishing the national database is seen in the examples of national databases described by 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.

The term national address database or dictionary (NAD) is sometimes used to refer both to any address database that claims to have national coverage (regardless of the data provider), as well as to an officially regulated register of addresses. To avoid confusion, in this chapter we refer to an official register of addresses as a national address register (NAR), and we use the term national address database (NAD) to include any national address database whether it is an officially regulated database or not.

6.2.2 The need for address data

Spatial address databases at all levels of government are required for ensuring services to a country's citizens. In South Africa, for example, according to the Bill of Rights in the constitution every citizen has the right to have access to, among others, adequate housing, a basic education, health care services, sufficient food & water, and social security. The constitution further stipulates how the different levels of government should ensure that these rights are delivered. However, a critical part of being able to deliver, for example, running water to citizens, is to know where the water has to be supplied. In the private sector there is also a need for a national spatial address database. As an example, South Africa's Financial Intelligence Centre Act (FICA) was written to assist in the identification of the proceeds of unlawful activities and the combating of money laundering. For that reason, customers of financial services institutions must provide proof of their residential address before opening an account. But how does a bank know that the address of a prospective customer is valid?

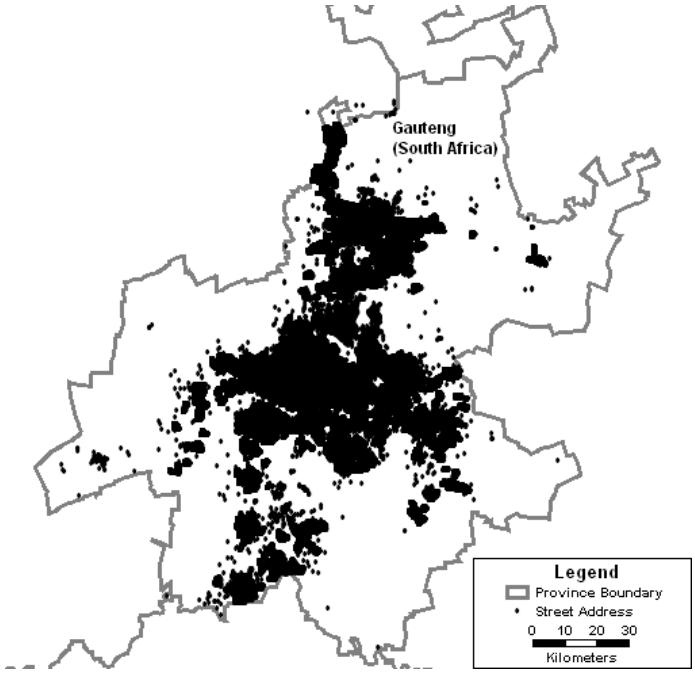


Figure 43. Street addresses in Gauteng (Source: AfriGIS NAD)

Other examples of address databases use are social services delivery where density of address data is used to prioritize the planning and roll-out of social services such as health clinics, schools and social service payout points in a country. Refer to Figure 43 for a map that shows the density of street addresses in Gauteng, a province of South Africa; goods delivery where courier, freight and logistics companies use spatial address databases to route their vehicles to a requested delivery

address; credit application where the residential address of the applicant is verified against a spatial address database; household surveys where the spatial address database is used for the delimitation of enumeration areas, as well as the planning and execution of surveys; elections for the delimitation of voting districts and the identification of voting stations in a country; emergency services to locate the emergency, and to route the relief team to the site (Yildirim and Yomralioglu 2004).

6.2.3 Spatial address data in South Africa

There is a large variety of address types in use in South Africa, as reported by Matheri (2005) and can also be seen from the draft South African address standard (SANS1883), which caters for street addresses, building addresses, farm addresses, informal addresses, intersection addresses, landmark addresses, various forms of postal addresses and site addresses (Coetzee and Cooper 2007b). The address type most commonly in use, is the street address type for which we have listed the Backus Naur form (BNF) in Figure 44. The map in Figure 3 shows a typical street address in a suburb in South Africa.

```
StreetAddress     = StreetAddressPart, Locality

StreetAddressPart = [CompleteStreetNumber | StreetNumberRange],
                          CompleteStreetName

Locality = RegisteredName | ColloquialName,
                          [TownName], [MunicipalityName],
                          [Province], [SAPOPostcode],
                          [Country] | [CountryCode]
```

Figure 44. The elements of a South African street address (SABS 2008).

In formal areas the StreetAddressPart is usually assigned by the municipality, but in informal areas and squatter camps this part of the address is randomly assigned. There is also the history of apartheid era townships in South Africa where only street names and no street numbers were assigned.

The Locality part of the address has one mandatory item: either the name of the suburb as recorded at a Surveyor General's office, or the name that is used colloquially for the area. The fact that people use both registered names and colloquial names results in ambiguity (and controversy) in names as used by the Surveyor General's office, municipalities, the SA post office and the general

public. For example, refer to 29 Queens Way in Figure 45. Because of the ambiguity in suburb names, an incoming address verification request for '29 Queens Way Hillcrest' could refer to any of the suburbs named 'Hillcrest' in Durban, Pretoria, Benoni, Kimberley, Wellington, Mthatha or Cape Town, of which only the suburb name 'Hillcrest' in Durban and Pretoria has been officially recorded at a Surveyor General's office. Further, since there is ambiguity in suburb boundaries '29 Queens Way' might actually be in Hadison Park, the suburb adjacent to the suburb named Hillcrest.

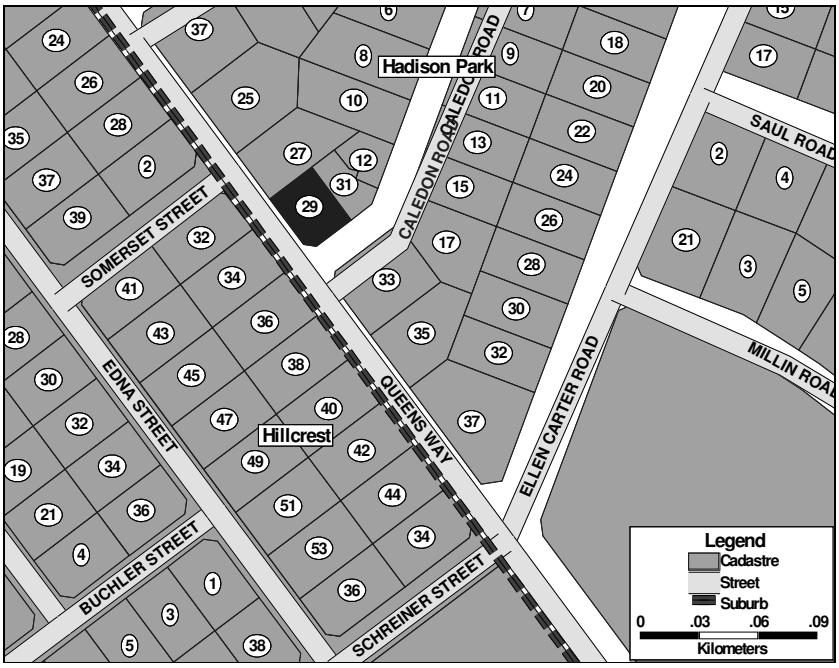


Figure 45. Hillcrest and Hadison Park in Kimberley (Source: AfriGIS NAD)

In 2001 South Africa was re-demarcated into 262 municipalities, and since then South Africa has been governed according to these municipal boundaries. However, people still use the 'town' names referring to the pre-2001 town councils in addresses. For example, the Akasia, Centurion and Pretoria town councils together with some other pre-2001 rural councils have been integrated to form the City of Tshwane metropolitan municipality. The names and boundaries of provinces and municipalities are determined and legalized by the Municipal Demarcation Board. Thus there is no ambiguity for the MunicipalityName and Province.

There are various sources of address data in South Africa, and some of these are listed in Table 23. The list is not comprehensive, but it illustrates that while there is not a single national address database in South Africa, there are a number of producers of address data that can each contribute to a national database of addresses.

The South African Spatial Data Infrastructure Act of 2003 was finally enacted in 2006, and the appointment of the Committee for Spatial Information (CSI) is currently (still) in progress. The act states that the CSI will appoint data custodians for SDI datasets. Thus, at the moment there is not a government appointed custodian for address data, and all the issues relating to custodianship are still open and have to be debated before any decision on custodianship is taken. It is therefore expected that custodianship will not be decided soon.

Table 23. Address data producers in South Africa

Source	Type of data	Purpose	Typical Coverage	Formats
GIS 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	A list of SA post office approved place names with their postcodes, no spatial information included	Postal mail delivery	National	Comma delimited text file
Statistics South Africa	Database of dwelling locations, address not always included	Household surveys	Per area as required for a survey	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 it into a national database	Address-related service provision, either by the company itself or sold to a third party	National	GIS database formats

Due to the current lack of a single government initiative to create a definitive national address database or register for public use, private organizations have identified and leveraged the business benefit of providing address-related products and services. These organizations source the address data for their national address databases from the sources listed in Table 2 and collate the data into a

national address database. The privately owned national address databases are distributed on a quarterly basis to clients in a single file in various formats. Clients of the national address databases include the private sector such as debt collectors, media companies, and financial institutions (banks and insurance companies) as well as the public sector such as SITA, Statistics South Africa, provincial and national departments of housing, and provincial and national transport authorities.

The cost of maintaining a national address database is high, and there are only a few organizations such as the major banks and large government organizations who can afford to buy the complete national address database. Private organizations have therefore started looking at new sources to recover some of the cost of data maintenance, and have started providing address-related services for which a user pays a small once-off fee for the service and use of data. For example, instead of paying hundreds of thousands of Rands for the national address database and then still having to implement an address verification service, the user pays R1 (approximately US$0.12 at the current exchange rate) or less (depending on volumes) to have a single address verified. Such a service makes the address data available to a much wider audience.

Regardless of how a national address data will be compiled for South Africa in the future – whether there will be one (or more) custodian(s) for address data, or whether a national initiative for a single national address database emerges, or whether address data will still be provided by private organizations – these address-related services are essential to making address data available to as wide an audience as possible. Based on this current scenario of address data in South Africa, we developed the evaluation framework that is described in the following section.

6.3 Evaluation framework

In this section we describe the framework that we use to evaluate potential information federation models for a national address database (NAD) in the South African context. Our chapter provides a technical evaluation of the models for a national address database, regardless of whether the national address database is officially regulated or not. To facilitate the evaluation, we present an architecture of conceptual layers for our national address database. Figure 46 illustrates these layers. In this section we describe the purpose of each layer, and then list the criteria of our framework by reference to the layered architecture.

The criteria of the framework are based on the requirements for the establishment, maintenance and use of a national address database and are summarized in Tables 24-30. The data provider layer contains the databases from the various address data providers. The unified view layer provides one or more common interfaces to any third party wanting to access the national address database. It also provides a unified view of the national address database, thus creating the illusion of working with a single database. In the service provider layer vendors provide services against the national address

database. Examples of services are an address verification service, an address geocoding service, or a mapping service. The application layer represents any application that makes use of a vendor service, for example, a home loan application form at a bank that makes use of an address verification service.

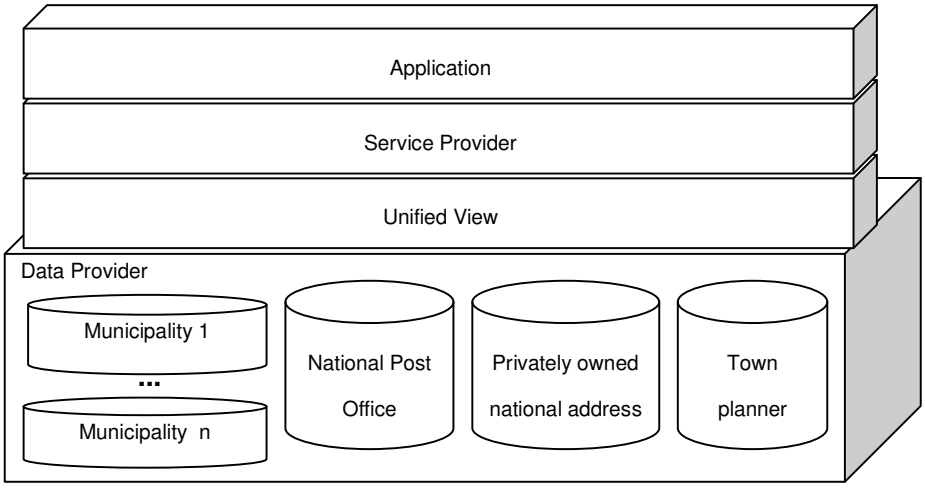


Figure 46. National address database

The first three criteria in our evaluation framework address heterogeneity in infrastructure (Table 24), data providers (Table 25) and naming conventions (Table 26). The following three criteria, namely address dynamics (Table 27), accessibility (Table 28) and security (Table 29), focus on issues around making the address data available to as wide an audience as possible. The final criterion addresses organizational issues (Table 30) of coordinating a national address database.

Table 24. Infrastructure

Criteria	Description
Operating system	Data and service providers should be free to use the operating system of choice.
Database Management System (DBMS)	A data provider should be free to store the address data in a DBMS (Oracle, SQL Server, ArcSDE, ESRI SHP files, MapInfo files, etc.) of choice.
Address format	Although address-related services should be based on a standardized address format, the unified view layer should accommodate the differences in address representation of the individual data providers.

Table 25. Data providers

Criteria	Description
Coverage area	Variation in the size and location of the coverage of address databases supplied by data providers should be allowed, and data access should be optimized for this, i.e. don't search for a Cape Town address in the Johannesburg database.
Decentralized source of data	The reality of many decentralized sources of address data providers must be catered for.
Multiple data providers per area	A data request should consider addresses from all the data providers, and resolve duplicates, ambiguities and potential semantic differences.

Table 26. Naming

Criteria	Description
Suburb Names	Enough information (such as alias tables) as well as disambiguating functionality should be provided to resolve between new official and colloquial names for suburbs.
Name Changes	Enough information (such as alias tables) as well as disambiguating functionality should be provided to resolve between new and old names of suburbs and streets.

Table 27. Address Dynamics

Criteria	Description
New developments	Address data for newly developed areas should become available as soon as possible. A quarterly update cycle is too long.
Previously un-addressed	Newly assigned addresses in previously unaddressed areas should be accessible as soon as possible in order to speed up service delivery to the areas as part of the development initiative in a country.
Address cross checking	Data providers should be able to cross check the availability of address data in areas for which they plan to produce address data.
Feedback from users to data providers	Users of the address data should be able to provide feedback to data providers about the correctness and accuracy of address data.

Table 28. Accessibility

Criteria	Description
Providing services (service providers)	Service providers should be able to provide value-adding address-related services on top of the unified view of the national address data. These services should be provided in a standard and well-known framework such as web services, and more specifically web feature services as specified by the Open Geospatial Consortium (OGC).
Billing and Accounting	The information federation model should allow a two-level billing and accounting system for both data use, and the use of vendor-supplied services.
Using services (application developers)	Application developers should be able to seamlessly integrate into their applications both services that provide access to the unified view of the national address database as well as the vendor-supplied services.
Access anytime	Access through these services to the national address database should be instantaneous and available all the time.
Access from anywhere	Access to the national address database should be available from as many platforms as possible including client desktops, personal digital assistants (PDA) and/or mobile phones.
Ease of publishing data (providing data)	Facilities for publishing address data should be easy and should not require specialized IT support.

Table 29. Security

Criteria	Description
User Authentication	Access to the national address database should be restricted to authenticated users.
Access	Data providers should be able to specify how and to whom (which group of people) their data is available.
Privacy	The data in the national address database should be protected against unauthorized access.

Table 30. Organizational Issues

Criteria	Description
Official custodians and unofficial data providers	The information federation model for a national address database should support the fact that there could be both officially regulated address data providers, supporting an official national address register, and unofficial address data providers, supporting national address databases in general.

6.4 Information federation models for a national address database

In this section we describe three distributed information federation models, namely data harvesting, federated databases and data grids. The models are commonly used for the federation of information but each has its own distinctive characteristics making it suitable for specific circumstances. We provide a description for each model, describe its purpose, and give examples of

its implementation. In order to further analyze the models, we list the sequence of events for performing a search service in each of the models. We describe each model by dividing it into four layers: application, search service, unified view, and the distributed data themselves, as illustrated in Figure 47. These layers correspond to the application, service provider, unified view and data provider layers in our conceptual architecture of a national address database. The difference between the models mainly lies in the way the data is stored and how the unified view of the distributed databases is achieved and maintained.

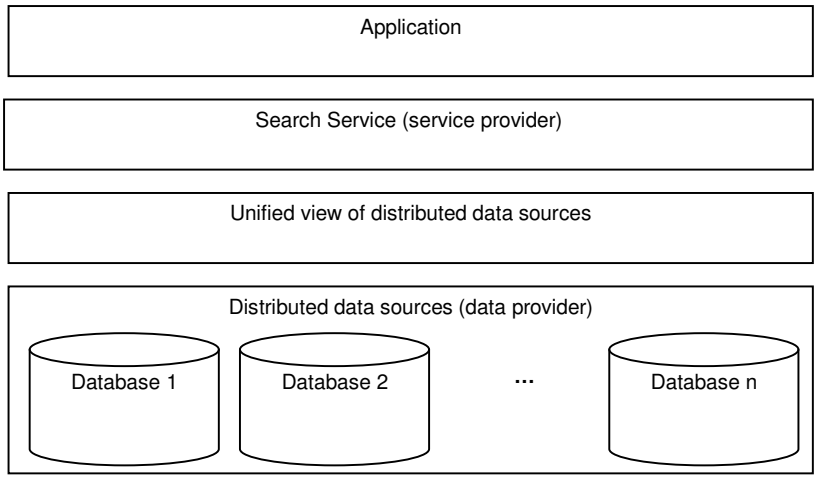


Figure 47. Information federation models

6.4.1 Data harvesting

In this model, data from a number of distributed databases is regularly harvested into a single centralized database, sometimes also referred to as data warehousing. Any search service accesses the single centralized database only, and does not have access to the distributed databases. The harvesting of data is either done online, e.g. through a web service, that pulls the data from one of the distributed databases and imports it into the centralized database; or harvesting is done offline by exporting the data from the distributed database and importing it into the centralized database. The underlying heterogeneity of the distributed databases, such as syntactic and semantic differences, is resolved when the data is harvested.

The centralized database is managed by a single organization, whereas the distributed databases are owned and managed independently. As long as one can export data into a format that can be imported into the centralized database, the management of the data in the distributed database is up to its owners.

The centralized database could be a relational database, but just as well a spatial or object-oriented database. The format (relational, spatial or object-oriented) of the individual distributed databases is also independent from the format of the centralized database. Data warehouse support provided by database management software such as Oracle, SQLServer or MySQL can be used to implement a centralized database.

Data queries are processed and optimized by the database management system (DBMS) that is used for the centralized database, but updates to individual data records are not possible as there is a uni-directional flow of data from the distributed databases into the centralized database. A centralized database has the potential of becoming a bottleneck but these can be resolved by load balancing techniques such as replication or mirroring of the centralized database. Since the centralized database is mostly read-only with regular and very specific types of updates, load balancing is easy to implement.

Figure 48 shows the sequence of events when performing a search for data in the data harvesting model. The dotted arrows indicate flow of harvested data.

1. The application calls the search service.
2. The search service queries the centralized database.
3. The resulting data is passed back to the search service.
4. The search service passes the resulting data back to the application.

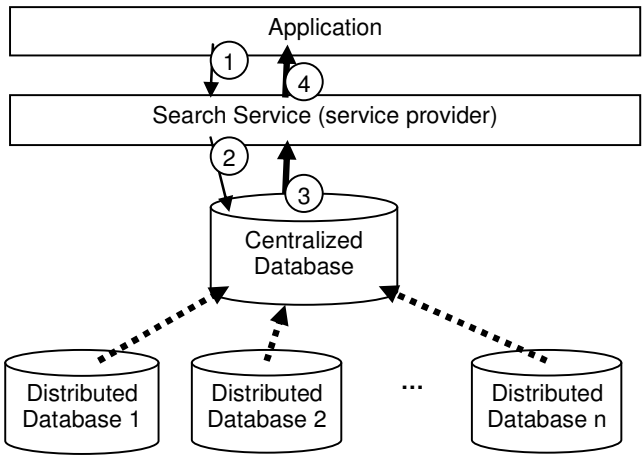


Figure 48. The data harvesting model

6.4.2 Federated database

A federated database (FDBS) is a collection of cooperating but autonomous component database systems (Sheth and Larson 1990). A significant aspect of a component database is the fact that it can continue with its local operations while at the same time participating in the federation. Federated databases are used to integrate existing diverse databases to provide a uniform, consistent interface for querying the underlying databases, and are sometimes also referred to as enterprise information integration. Federated databases accommodate any kind of underlying heterogeneity in terms of representation and syntax in the component databases. Federated databases are tightly integrated systems and usually maintained by a single organization.

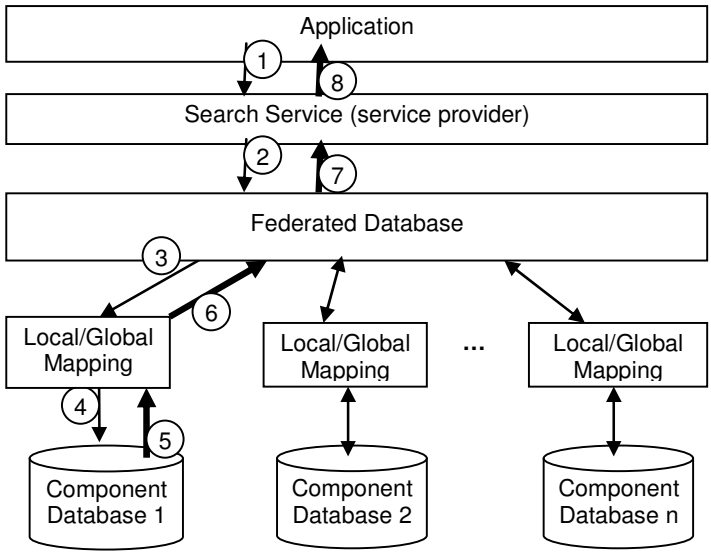


Figure 49. The federated database model

A database management interface provides access to the FDBS, and data records are both read and written frequently, thus necessitating transactions. Some form of query language, such as SQL for relational databases, is used to construct queries. The FDBS interprets, optimizes and executes the queries against the underlying component databases and provides results back to the querying process. The federation is established by mapping the local representation of a component database to the global representation of the federated database. The purpose of an FDBS is to integrate existing heterogeneous databases and to provide a uniform and consistent interface for querying and updating data in the underlying databases.

Figure 49 shows the sequence of events when performing a search for data in the federated database model. The thick arrows indicate data flow.

1. The application calls the search service.

2. The search service queries the federated database.
3. The query is translated into a form that the component database understands, i.e. there is a translation from global to local representation and syntax. Semantic differences, as well as data schema differences, in the underlying component databases are resolved.
4. The query arrives at the component database and is executed.
5. The resulting data is mapped from local to global representation and syntax. Semantic and data scheme differences are resolved.
6. The resulting data (global view) is passed back to the federated database.
7. The federated database passes the resulting data back to the search service.
8. The resulting data is passed back to the application.

The concept of a federated database has been applied to georeferenced data where existing spatial databases are integrated into a single map view with a uniform, consistent interface for querying, navigating and/or updating the underlying spatial databases. Earlier work by Coetzee and Bishop (1998) presented the design and implementation for a distributed open spatial query mechanism in Java, using Java Native Interface (JNI) and Remote Method Invocation (RMI) that provided a uniform view to heterogeneous spatial data sources. Tuladhar *et al.* (2005) propose a federated data model for distributed cadastral databases for land administration in Egypt. Another example would be a single map generated at a local authority that displays land parcel boundaries from an ArcSDE database in the town planning department and street centre line data from an Oracle spatial database in the engineering department. IBM's Information Integrator together with the IBM WebSphere Federation Server (refer to www.ibm.com), give real-time access to distributed databases in such diverse formats as Oracle databases, Microsoft Excel spreadsheets and flat files. A consistent view of data is created and federated access to the multiple data sources is provided.

6.4.3 Data grid

The term 'grid' has been used in many ways, including everything from advanced networking to artificial intelligence. To eliminate confusion, in our discussion we stick to the definition of a grid as defined by the Open Grid Forum (Treadwell 2006): 'A system that is concerned with the integration, virtualization, and management of services and resources in a distributed, heterogeneous environment that supports collections of users and resources (virtual organizations) across traditional administrative and organizational domains (real organizations).' We thus exclude cluster computing or so called computing on demand, which is provided and marketed as 'grid' by some of the commercial companies, including Oracle.

A data grid is a specific type of grid where the resources are databases or data files. A data grid provides services that help users discover, transfer, and manipulate large datasets stored in

distributed repositories and also, create and manage copies of these datasets. Data in a grid is syntactically, structurally and semantically heterogeneous but the grid provides an integrated view of data, which abstracts out the underlying complexity behind a simple interface. The word 'grid' is an analogy with the electric power grid, which provides pervasive access to electric power (Foster and Kesselman 1999). Similarly, the idea behind a data grid is to provide pervasive access to data.

In a data grid, each participating node has full autonomy in terms of operations (the node conducts its own operations without being overridden by external operations), participation (the node can decide on the proportion of its resources to be shared in the grid), and access (the node can decide to whom access should be granted). Data grids are mostly read-only environments into which existing data is introduced or replicated. If the source of a data replica is updated, its corresponding replica on the grid is also modified (Venugopal 2006). Currently data grids do not provide support for transactions, but the topic is on the agenda of the Open Grid Forum (OGF Transaction Management Research Group 2005).

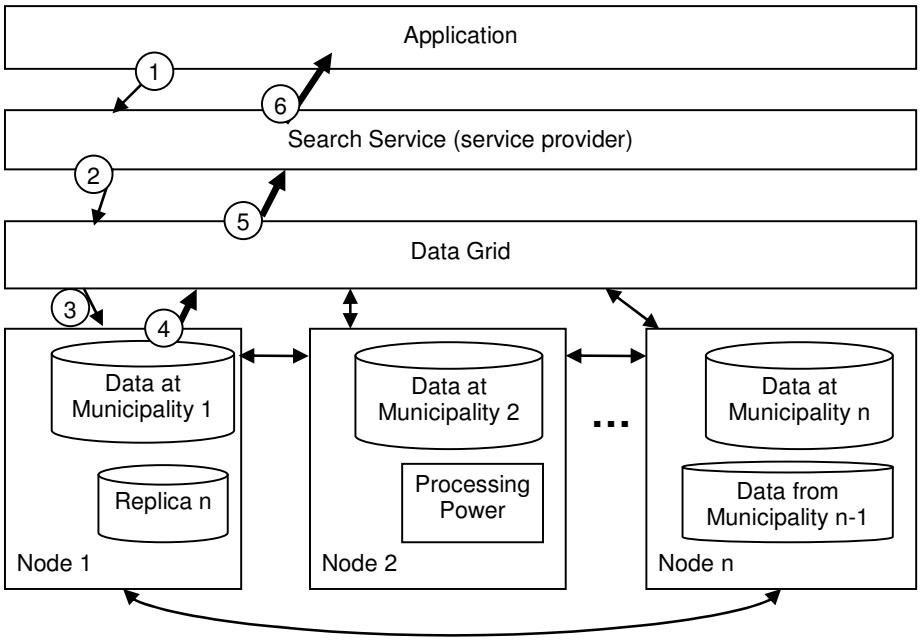


Figure 50. The data grid model

Data grids carry metadata about the collaborating datasets, which is stored in a metadata catalogue and carries the logical dataset name together with the physical locations of the dataset and its replicas. The metadata can also include other attributes, such as those specified in ISO 19115, *Geographic information–Metadata*, to describe the data, which can then be included in any data query.

The Open Grid Services Architecture – Data Access and Integration (OGSA-DAI) is a

middleware product which supports the exposure of data resources, such as relational or XML databases, onto grids. Consistent interfaces to a number of popular database management systems are provided, and a collection of components for querying, transforming and delivering data via web services is also included. (OGSA-DAI website 2008).

Figure 50 shows the sequence of events when performing a search for data in a data grid. The thick arrows indicate data flow.

1. The application calls the search service.
2. The search service queries the data grid.
3. The data grid locates the correct replica and does the necessary translations. It then passes the query to the node with a current replica of the data.
4. The resulting data is passed back to the data grid.
5. The data grid does the necessary backward translations and passes the resulting data back to the search service.
6. The resulting data is passed back to the application.

The Globus Toolkit, an open source software toolkit for building grid systems and applications, is developed by the Globus Alliance, an international collaboration that conducts research and development to create fundamental grid technologies. Its members include the Argonne National Laboratory at the University of Chicago, the National Center for Supercomputing Applications (NCSA) in the US, Univa Corporation, the University of Southern California Information Sciences Institute and the Royal Institute of Technology in Sweden.

On the commercial front Sybase Avaki Data Grid (refer to www.sybase.com) is a commercially available data grid solution where data remains with the authoritative sources, thereby eliminating inconsistencies and complexities introduced in managing multiple copies of the data required for compute grid applications. Avaki handles the performance and scalability needs in a clustered grid, an enterprise-wide grid, or across a grid spanning multiple administrative domains.

Examples of data grids in the earth sciences that are based on georeferenced data are the Earth Systems Grid which integrates peta-bytes of data with analysis resources to provide an environment for next generation climate modeling and research; and NEESgrid which is used by earthquake researchers to aggregate information from sensor equipment, and used on a platform of high performance computing to design and execute experiments. The modeling and simulation of biological processes, coupled with the need for accessing existing databases, has led to the adoption of data grid solutions in the bio-informatics discipline. These projects involve federating existing databases and providing common data formats for the information exchange (Venugopal 2006).

6.4.4 Comparative analysis

Table 31 below provides a comparative overview of the three information federation models presented in this section.

Table 31. Comparative analysis of information federation models

	Data harvesting	Federated database	Data grid
Purpose	Aggregate data from diverse databases into a single centralized database	Provide an integrated view on existing diverse databases with a uniform and consistent interface	Provide services to discover, transfer, and manipulate large datasets stored in distributed databases and giving an integrated view of the data
Unified view provided by	Single centralized database of data	Uniform and consistent interface to the federated database	Standardized data grid services
Syntactic translation and semantic interpretation	Once off when harvested data is loaded into the centralized database	With each access	With each access
Data updates	No, read-only	Equally read and write	Mostly read with rare writes
Transaction support	Read-only access does not require transactions	Yes	Not yet (being researched)
Architecture	Service-orientation for access to the centralized database	Service-orientation for unified data access	Service-orientation for unified data access and underlying architecture

6.5 Evaluation

In this section we describe the implementation issues for each model in the context of a national address database, and go on to analyze such an implementation based on the criteria set out in our evaluation framework for a national address database in South Africa. A comparative analysis is provided at the end of the section.

6.5.1 Single centralized harvested national address database

Figure 51 illustrates a national address database that is harvested from a number of data providers. We have added the four layers from our evaluation framework as a reference in the figure. Address data from the data providers is harvested at regular intervals and loaded into the single centralized database.

An additional layer of abstraction on top of the central database provides standardized technology-independent access to the database, and we call this layer the standardized NAD services. Once again, the OGC Web Feature Services are a suitable specification for services that query and retrieve address data from the central database. These standardized NAD services provide

access to the centralized database in a uniform way with the fundamental services required such as traversing through the NAD in a specific suburb, finding a specific address record, etc. Application developers either access the central NAD through the standardized NAD services, or use the specialized services provided by independent service providers.

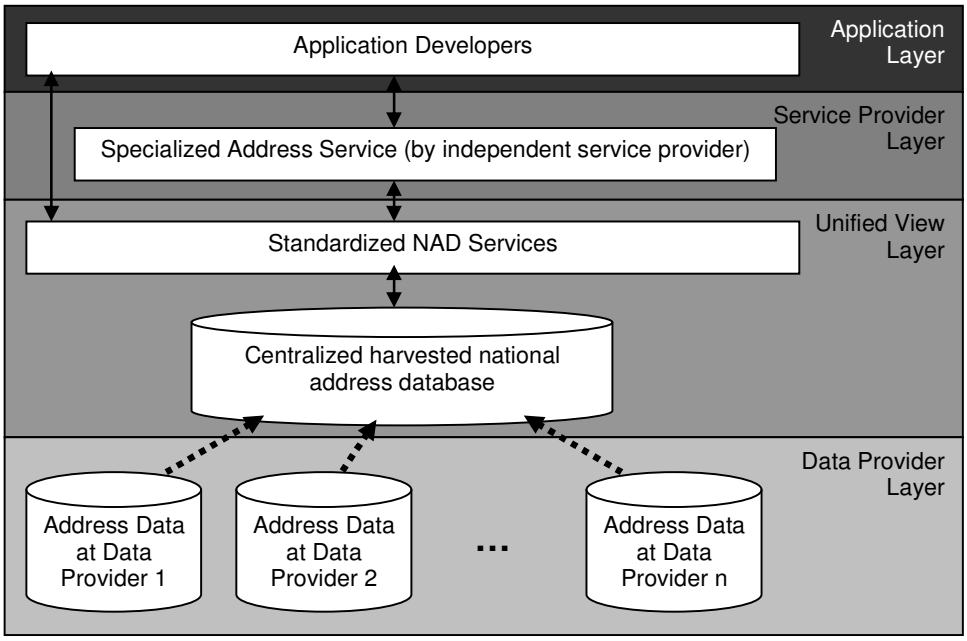


Figure 51. Single centralized harvested national address database

6.5.1.1 Examples

Australia. The Australian Geocoded National Address File (G-NAF®) is updated in an incremental format quarterly – usually in February, May, August and November. The Public Sector Mapping Agencies (PSMA) follows a semi-automated process of massaging contributor address data into a standardized format that is acceptable for merging into the G-NAF. Any address data that cannot automatically be converted into the standard address format, is subjected to a manual review process. The data is distributed in a format known as a MapInfo file (GIS) in a single GIS data file. The PSMA is the custodian of the Geocoded National Address File (G-NAF). However, they are not the source of the data; PSMA acts as a clearinghouse by merging data from as many as 15 government agencies and organizations into the G-NAF (Paull 2003).

Ireland. In Ireland a definitive reference directory for addresses is maintained by An Post and Ordnance Survey Ireland (OSi). The GeoDirectory, as it is called, combines postal addresses (where mail is delivered) and geographic addresses (a geo-code to position the address on a map) in one database, which is available to organizations or individuals who require it. GeoDirectory updates are

released four times a year by supplying customers with a single completely refreshed database (Fahey and Finch 2006).

6.5.1.2 Evaluation

Infrastructure. The standardized NAD services and/or the data exchange format of address data files accommodate heterogeneity in terms of operating system, DBMS and address data format. Other heterogeneity is eliminated when the data is loaded into the single centralized database.

Data Providers. Different coverage areas of individual datasets are irrelevant in the data harvesting model, as all data is loaded into a single database. Duplicate addresses as provided by multiple data providers are either resolved when loading the data into the centralized database by applying a set of rules for picking the most pristine address to be loaded; alternatively duplicate addresses are loaded into the single database and the user specifies with parameters to each address data request which address data should be included in the query. Example parameters are a specific data provider, and minimum accuracy and quality requirements.

The data harvesting model accommodates the decentralized sources of address data by aggregating it into a single centralized database. However, a data provider gives up some of its autonomy by handing over the data to a centralized database. There is now a middle party – the administrator(s) of the centralized database.

Naming. A table of old and new names of places, as well as official and colloquial suburb names is stored in the single database. The table should include a spatial boundary for each name so that addresses such as the '29 Queens Way Hillcrest' problem described earlier can be resolved by searching surrounding suburbs. Any request for address data uses these tables to disambiguate a request for address data.

Address dynamics. In the data harvesting model the currency of the address data depends on how fast new and modified addresses can be loaded into the centralized database. From the Australian example it is clear that this process, even in a regulated environment, can be quite tedious involving manual reviewing of data.

In order to prevent duplication of efforts, data providers use the standardized NAD services to cross check whether an address already exists. Since all data is in one single database, summarized reports of address data per area can be published.

The feedback cycle from the general public involves three parties: the person in the general public who generates feedback to the provider, the data provider who modifies the address data if required, and the centralized database into which the modified address is loaded.

Accessibility. The standardized NAD services provide platform independent access to the

address data to both application developers and service providers. Access anytime and from anywhere is addressed by providing online access to the single database via the standardized NAD services. The responsibility for up time lies with the single entity in charge of the centralized database. For better performance, the single database can be replicated and load-balancing techniques applied.

A potential problem in the model followed in the Australian and Irish examples above is that copies of the single centralized database are distributed to buyers of the data. Online access to the data is not the aggregator's responsibility, but that of whoever purchases the database and provides online access to it. This could result in a situation where service provider A makes services available on its copy of the database from the first quarter of a year, while service provider B's services are available on its copy of the database from the third quarter of a year. To an application developer who uses services from service providers A and B this results in conflicting views of the address data.

In the single database environment, billing for address data is handled by any of the current online transaction environments. Billing models include paying for accessing specific address data or paying a monthly subscription fee. Billing and accounting for use of the specialized services should be done by each independent service provider.

Security. In the case of the data harvesting model, security measures such as user authentication and granting access to data is implemented by the centralized database. Most database management systems, whether relational, spatial or object-oriented, have support for these security measures.

Organizational Issues. The data harvesting model requires a single organization to control and administrate the centralized national address database. If there is no organization with the mandate or the financial means to do this, the implementation of the data harvesting model is difficult, as it is preferable that some organization take responsibility for the coordination and loading of address data into the single centralized database.

6.5.2 A federated national address database

In this model each data provider makes its database of address data available to the federation. A data provider's database has to be online in order to participate in the federated national address database, but it can be used for any other local operations while participating in the federation. Figure 52 illustrates the mapping between local and global representations in the architecture of a federated national address database.

The address data specific mappings, such as interpreting semantic differences, are implementation dependent and have to be developed specifically as part of the federated national

address database. The unified view layer exposes a set of standardized NAD services, similarly to the harvested NAD.

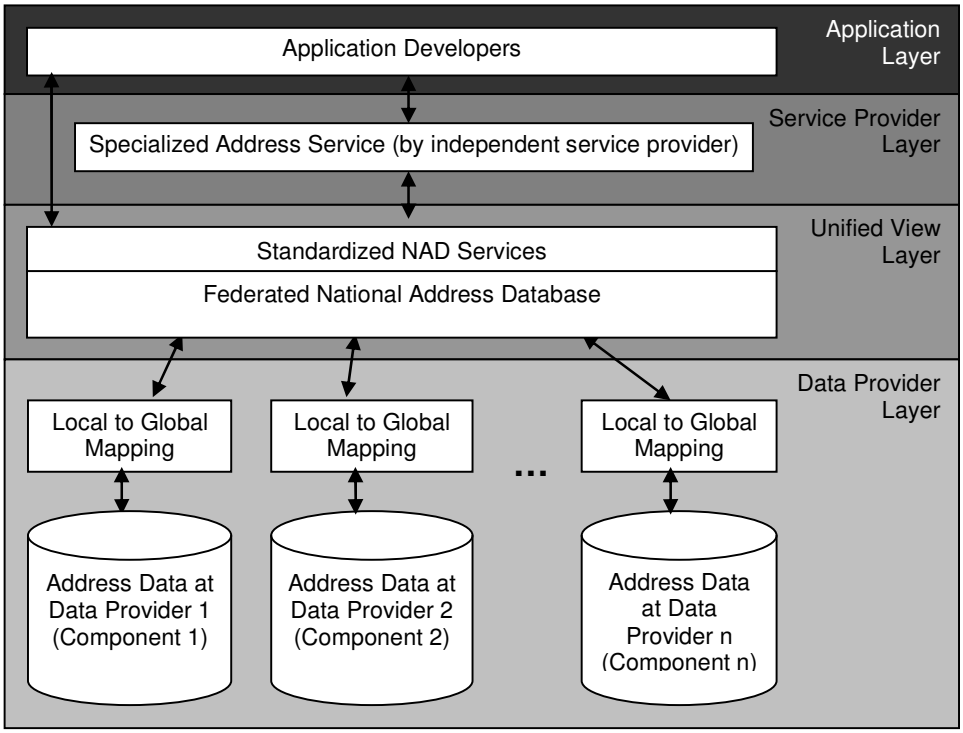


Figure 52. Federated national address database

6.5.2.1 Examples

Egypt. Although the example we present here is not a federated national address database, it is an example of a federated national database of land information, in many ways similar to address data. Tuladhar *et al.* (2005) propose a federated data model for the situation in Egypt where land ownership, state owned land data, cadastral data, topographic data and tax data are maintained by four different government departments. These datasets are maintained and stored at their respective departments at provincial level (i.e. sub-national level). The federated data model allows integrated access to the databases on a national level, while control over the maintenance of the data remains at the provincial government departments.

6.5.2.2 Evaluation

Infrastructure. In the federated database mapping from local to global data representation happens on the fly with each data request, thus the complexity of the local/global mapping

influences the performance of address data queries.

Data Providers. The federated database by definition provides access to decentralized sources of data. Metadata such as the coverage area of a dataset and the data provider for the dataset are stored in separate tables (either at individual data providers or at a centralized location) and used whenever a distributed query is executed. Duplicate addresses from multiple data providers are either resolved by the distributed query mechanism, or passed back to the requester to resolve. For example, if the requester is an independent service provider, a statistical probability for the address with the largest probability of being correct can be added before passing the address back to the application layer.

Naming. The old and new names of places are stored for example, in a designated component database; the same applies to official and colloquial suburb names. The federated NAD cannot rely on underlying data providers to resolve all naming ambiguities; therefore the disambiguation functionality has to be implemented in the unified view layer.

Address dynamics. The currency of address data depends on the currency of the underlying component database. Since these databases reside with the data providers, there is no delay from updating to publishing address data. As soon as the data is updated in the component database, it is available in the federated NAD.

In order to prevent duplication of efforts, data providers can use the standardized NAD services to cross check whether an address already exists.

The feedback cycle from the general public involves two parties: the person in the general public who generates feedback to the provider, and the data provider who modifies the address data if required.

Accessibility. The standardized NAD services provide platform independent access to the address data, and can be used by both application developers and independent service providers. In the data harvesting model there is one entity – the centralized database – of which the uptime has to be managed; in the federated database each individual component database's uptime has to be ensured. If one of the components is off-line, the accessibility of the federated national address database is reduced, but the remaining parts of the federated database can still be accessed.

Billing for address data is handled by any of the current online transaction environments and has to be integrated into the federated database on the unified view layer. Billing and accounting for use of the specialized services should be done by each independent service provider.

Security. Security measures such as user authentication and granting access to data are implemented in the federated database as part of the unified layer. A user with access to an

underlying component database does not have access to the federated database, but a separate user account on the federated database level is required.

Organizational Issues. Federated databases are typically created within a single organization. The participation of a component database is granted and controlled from a central point. If there is not a single organization with the mandate to establish and maintain a national address database a tightly coupled solution such as a federated database is difficult to implement.

6.5.3 National address data grid

In the national address data grid, each data provider makes its address data available on the grid, and can opt to make other resources such as storage space and processing power available as well. Figure 53 illustrates the components involved in the data grid approach for a national address database. Since data grids are mostly read-only environments into which existing data is introduced or replicated, this fits the scenario of each local authority maintaining its own address database but making it available to the national address data grid whenever it is updated. Interoperability mechanisms to handle the heterogeneity in address format and semantics of the underlying data providers' databases has to be developed specifically for the national address data grid.

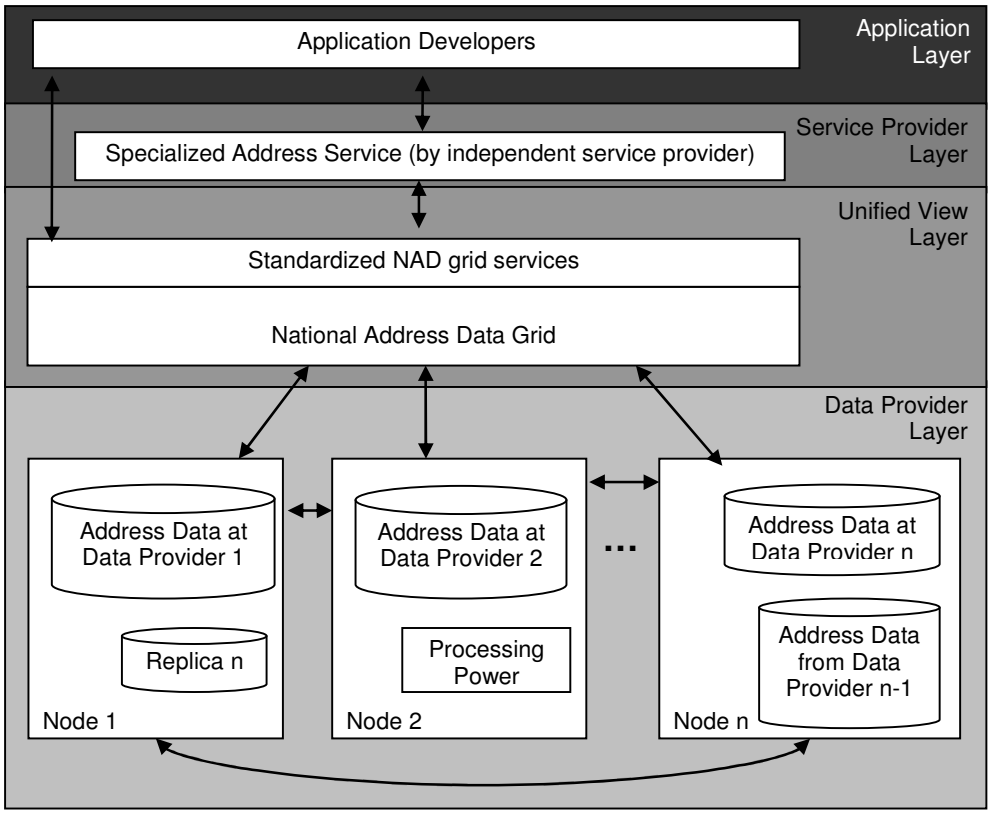


Figure 53. The national address database as a data grid

The standardized NAD grid services once again provide the uniform view to the underlying heterogeneous data sources. Venugopal *et al.* (2006) provide a taxonomy for data grids. According to this taxonomy, a national address data grid is organized as a federated model of stable data sources with inter-domain scope where the virtual organization is created for collaboration and economic benefit of the individual participants and possibly regulated by a national authority at a later stage.

6.5.3.1 Evaluation

Infrastructure. In the data grid model the grid middleware addresses operating system heterogeneity, and OGSA-DAI is an example of grid middleware that takes care of difference in individual data providers' data representation. OGSA-DAI is compliant with the Globus Toolkit and also entirely implemented as web services, therefore providing a platform independent solution.

Data Providers. The metadata catalogue stores information about the decentralized sources of data including the coverage area of a dataset. Duplicate addresses from multiple data providers are either resolved by the distributed query mechanism, or passed back to the requester to resolve. Similarly to the FDBS, if the requester is an independent service provider a statistical probability for the address with the largest probability of being correct can be added before passing the address back to the application layer.

Naming. Old and new names, as well as official and colloquial names can be stored in anyone of the decentralized data sources in the grid. Similar to the federated database, the national address data grid cannot rely on underlying data providers to disambiguate all names, and thus the disambiguation functionality has to be implemented in the unified view layer as part of the grid middleware.

Address dynamics. In the data grid model the currency of address data depends on the currency of the underlying data providers' databases: as soon as the data provider has updated its address data, it is available to users of the NAD services. There is no time delay from update to availability.

Similar to the other two models, data providers can use the standardized NAD services to cross check whether an address already exists in order to prevent duplication of efforts.

The feedback cycle from the general public involves two parties: the person in the general public who generates feedback to the provider, and the data provider who modifies the address data if required.

Accessibility. The standardized NAD services provide platform independent access to the address data, and can be used by both application developers and service providers. Access anytime and from anywhere is addressed by replicating the data provider databases in the grid; in the data

grid, the uptime of several core nodes has to be ensured (and not the uptime of each individual node).

Data billing and accounting information can be handled by the grid middleware. There is somewhat more complexity involved in this model when not only data but also computing resources are shared.

Security. Security measures such as user authentication and granting access to data are taken care of by grid middleware. The virtual organization model is applied whereby for example, a user's access rights to data are derived from his/her membership in the virtual organization. This makes authentication more complex than in the other two models, but it has the advantage that user accounts do not have to be created by a central authority. Since the grid paradigm is still relatively new, not all security issues have been addressed by the grid community yet. However there is a lot of current research in this area.

Organizational Issues. A data grid provides the required flexibility of data providers entering and leaving the scene of contribution to the national address database. Thus the data grid could survive the transition from a national address database to which both officially regulated and unofficial address data providers contribute, to a national address register to which only officially regulated address data providers contribute. The data grid also does not rely on a single central organization to control and administrate the national address database, but allows a more organic type of existence with multiple contributors.

Harvey and Tulloch (2006) describe the 'federation-by-accord' data sharing model, which involves a number of data producers who generally share their data with a number of other data users and producers in their network. The model is resilient to change and can afford to lose a major player without ruining the entire model. They found that this model approaches the ideal national SDI data sharing environment in many ways, and that if it is integrated into the ongoing activities of local authorities, it becomes sustainable and the vehicle for enhancing data sharing. A data grid would support such a 'federation-by-accord' data sharing model.

6.5.4 Comparative Analysis

Tables 32-38 provide a comparative overview between the three information federation models in relation to the criteria of our framework.

Table 32. Infrastructure

Criteria	Data Harvesting	Federated Database	Data Grid
Operating system	Once off when loading the data into the single centralized database	Dynamically with each data request	Dynamically with each data request
DBMS heterogeneity	Once off when loading the data into the single centralized database	Dynamically with each data request by middleware such as ODBC or JDBC	Dynamically with each data request by the grid middleware, e.g. OGSA-DAI
Address data format	Once off when loading the data into the single centralized database	Dynamically with each data request	Dynamically with each data request

Table 33. Data providers

Criteria	Data Harvesting	Federated Database	Data Grid
Coverage area	Irrelevant as all data is in one database	Stored in separate metadata tables	Stored in the metadata catalogue
Decentralized source of data	Not possible	Component databases	Grid nodes
Multiple data providers per area	Either when loading the data or stored as an attribute of the address	Resolved on the fly or passed back to the requester to resolve	Resolved on the fly or passed back to the requester to resolve

Table 34. Naming

Criteria	Data Harvesting	Federated Database	Data Grid
Suburb names and name changes	Disambiguation information stored in the centralized database Disambiguation functionality provided by the centralized database	Disambiguation information stored in one of the component databases Disambiguation functionality provided by the federated database	Disambiguation information stored at one of the grid nodes Disambiguation functionality provided by the data grid middleware

Table 35. Address dynamics

Criteria	Data Harvesting	Federated Database	Data Grid
New developments	Time delay	Immediate	Immediate
Previously unaddressed areas	Time delay	Immediate	Immediate
Address cross checking	Standardized NAD services	Standardized NAD services	Standardized NAD services
Feedback	Three parties	Two parties	Two parties

Table 36. Accessibility

Criteria	Data Harvesting	Federated Database	Data Grid
Providing services	Platform independent web services such as OGC web feature services	Platform independent web services such as OGC web feature services	Platform independent web services such as OGC web feature services
Billing and accounting	Online transaction environment	Online transaction environment	Still being researched
Using services	Platform independent web services such as OGC web feature services	Platform independent web services such as OGC web feature services	Platform independent web services such as OGC web feature services
Access anytime	Single server	Each server with a component database	A number of core nodes
Access from anywhere	Internet	Internet	Internet
Ease of publishing	Data providers have to convert their data into the address data exchange format	Data providers store data in their choice of database	Data providers store data in their choice of database

Table 37. Security

Criteria	Data Harvesting	Federated Database	Data Grid
User authentication, access and privacy	User accounts in the centralized database Data updates and transactions not possible	User accounts of the federated database Data updates and transactions are allowed in the federated database, but should be controlled by the local data provider for proper dataset management	Authentication is established through the virtual organization Data updates are theoretically possible, but transactions not yet available

Table 38. Organizational Issues

Criteria	Data Harvesting	Federated Database	Data Grid
Official custodians and unofficial data providers	Requires central coordination and organization	Requires central coordination and organization	Provides flexibility for data providers to come and go

6.6 Conclusion

We have presented the status of spatial address data within the context of SDI and have thereby illustrated that the sources for address data are distributed and not under centralized coordinated control. We illustrated the need for address data in both the public and private sector, and justified the need for address-related services on a national level, making specific reference to South Africa. Thus, there is a demand for non-trivial address-related services. We have further shown that there are typically numerous and diverse sources of address data, resulting in ambiguities and heterogeneities in the address data. Therefore, one has to work with standard, open interfaces for address data content as well as access to the address data. These three features of address data are closely related to the three-point checklist for a grid provided by Foster (2002).

Our novel evaluation framework describes important criteria for a national address database and we use the South African scenario to contextualize the framework. We used this framework to evaluate three information federation models: data harvesting, federated databases and data grids, and compare implementation issues for a national address database in the form of each of the models. The large number of organizations involved in a national address database, as well as the lack of a single organization tasked with the management of a national address database, presents the data grid as an attractive alternative to the other two models. The data grid provides for a more loosely coupled architecture, thereby allowing for more diversity and heterogeneity.

The typology for local government sharing in the United States, as presented by Harvey and Tulloch (2006), describes some disadvantages to giving a single organization the authority over data production and sharing. Both the data harvesting model and the federated database model require a single organization to take control. Harvey and Tulloch report that a federation-by-accord, although difficult to establish, once integrated into ongoing activities, can become sustainable and a suitable vehicle for enhancing data sharing. Our novel approach to a national address database as a data grid corresponds to the ‘federation-by-accord’ data sharing model which can afford to lose a major player without ruining the entire model.

As part of our THRIP project, which is funded by the Department of Trade and Industry (dti) and our industry partner, AfriGIS, we are setting up a data grid with the Globus toolkit at the University of Pretoria, and are busy expanding it to AfriGIS and our collaborators on the project in Dhaka, Bangladesh. Some very basic address verification services are currently running on the grid at the university, and the plans are to expand on these. As part of our research we are currently investigating charging frameworks for a national address database on the grid. Our data grid benefits from the service-oriented architecture of the Globus Toolkit, which provides for a loosely coupled solution. We believe that there are also large benefits to be gained from the more traditional grid services in Globus such as those for resource scheduling (GRAM) and large file transfers (GridFTP),

and this provides for interesting research questions for future phases of our research.

Data grids are a more recent development and current implementations are still mostly in the scientific research environment. At this stage most data grid implementations focus on high volumes of data and high processing loads whereas an implementation of a national address data grid would focus on pervasive access to address-related resources (data and services), as envisaged with the original analogy to the electrical power grid.

Chapter 7 Conclusion

7.1 Introduction

The work in this dissertation was a first investigation into the data grid approach to national address databases in an SDI, which has led to more research questions to be addressed by future research. This final chapter of the dissertation provides both a retrospective view on the main results from the work in this dissertation, as well as an outlook to the future.

7.2 Main results from this dissertation

This dissertation presents an analysis of the data grid approach for spatial data infrastructures. The two imaginary scenarios that were devised by the author and presented in Chapter 1, for the first time spell out how data grids can be applied to enable the sharing of address data in an SDI, so that services can be realized that are beyond the capacity of an individual organization. The novel evaluation framework for national address databases is used to evaluate existing information federation models, as well as the data grid approach, for the use in address databases for national SDL. This evaluation, as well as an analysis of address data in an SDI, confirms that there are quite a few similarities between the data grid approach and the requirement for consolidated address data in an SDI. The evaluation further shows that where a large number of organizations are involved, such as for a national address database, and where there is a lack of a single organization tasked with the management of a national address database, the data grid is an attractive alternative to other models.

Currently, most national address databases of the world follow the centralized approach where address data is loaded into a single centralized server. The novel data grid approach proposed in this dissertation deviates from this centralized approach, and is therefore different to current approaches. While there are research projects investigating the sharing of geospatial data on a grid, the work described in this dissertation focused on the specific case of address data in an SDI.

Today, still, address data is considered as a mere attribute of an entity in many a corporate system, instead of being regarded as a reference. The ‘address as an attribute’ notion does not require validation of the combined address fields because the address is stored as a number of separate text attributes; whereas an ‘address as a reference’ ensures that the address exists in a reference dataset and that any changes to the referenced address are automatically linked back to the entity that refers to it. The ‘address as an attribute’ notion is the source of invalid and ‘dirty’ address data in many a customer database (from personal experience of the author), resulting in problems further

downstream when, not only, for example, customers have to be geocoded for spatial analysis or routing, but also when postal mail has to be delivered to those customers. The definition for an address and an addressing system that were provided in Chapter 2 further enhance the understanding of what an address is. An address data grid, such as that proposed in the Compartimos reference model, enables wider access to an address reference dataset as part of an SDI, thus contributing to the accuracy and quality of address data in general.

The different formats and models used for address data at the various local authorities where the address data is produced and maintained, pose a major challenge to data integration in a grid environment. The interoperable address data model that is proposed as part of Compartimos in Chapter 4 is one way of enabling address data interoperability. If, for example, the interoperable address data model from Chapter 4, were adopted as an international address standard, an international address data grid would be feasible so that it is possible to present a single virtual address dataset of all address data in the world. This data model is based on ISO 19112, a standard published by ISO/TC 211, showing how an application domain-specific standard can be built on the foundation that has been established by application domain-generic standards. Data grid standardization takes place through the Open Grid Forum (OGF) in cooperation with OASIS, while standardization of geographic information and associated services takes place through the ISO/TC 211, *Geographic information/Geomatics* and the Open Geospatial Consortium (OGC). *ISO/TC211–Geographic information* recently voted in favor of projects that will develop an interoperable data model for land administration (cadastre and property rights) and the classification of climate change variables which can be seen as a sign that ISO/TC211 will in future expand its scope to include more work on application domain specific standards, including address data.

Compartimos has a clarifying purpose, because it is one of the first attempts in the joint SDI and data grid problem space, and thus enhances the mutual understanding between the geographic information community and the data grid community. This mutual understanding is one of the goals of the recently announced collaboration between the OGF (data grid problem space) and the OGC (SDI problem space). The initial focus of this MoU is to integrate OGC's OpenGIS Web Processing Service (WPS) Standard with a range of "back-end" processing environments to enable large-scale processing, or to use the WPS as a front-end interface to multiple grid infrastructures, such as TeraGrid, NAREGI, EGEE and the United Kingdom's National Grid Service. Research results from this dissertation suggest that there is also an opportunity for spatial data integration in an SDI environment that should be explored, in other words, a requirement to grid-enable other web services specified by OGC, such as, for example, the Web Feature Service (WFS).

In order to analyze and reason about the data grid approach to address data consolidation in an SDI, it was necessary to define ‘the animal’: Compartimos, a reference model for an address data

grid, was developed in order to get a better understanding of all the components involved in such a data grid approach. Compartimos is an abstract representation of the entities and relationships that realize such an address data grid in an SDI. Compartimos serves to analyze the problem space of data grids and SDIs by addressing a very specific problem in these areas. The discussion of the technology choices for Compartimos objects in Chapter 5 contributes towards the understanding of how far down the road we are in terms of developing an address data grid in an SDI. The discussion looks at how existing technology can be used, and in which areas research and development is still required. Compartimos is also a novel application of the OGSA data architecture, intended as a general architecture, in the (very specific) environment of address data in an SDI. The proof-of-concept implementation of Compartimos in a controlled environment represents a specific combination of technology choices. The results and recommendations that are drawn from the experience of designing and implementing Compartimos are valuable for future research in this area.

The future lies in distribution and integration, two seemingly contradicting nouns. Due to the Internet, wireless networking and mobile devices, it is possible to stay connected to the global network always and wherever you are – resulting in more distribution. As a result, there is an increase in the supply of diverse information that needs to be integrated and assimilated in order to be understood.

7.3 Recommendations for further research

The work in this dissertation was a first investigation into the viability of the data grid approach to national address databases in an SDI. The following sub-sections describe five issues that warrant further research:

1. Grid enabling OGC web services
2. Developing an international address standard
3. Trusting address data resources
4. Generalizing Compartimos for all kinds of spatial data in an SDI
5. SDI in the clouds

7.3.1 Grid-enabling OGC web services and spatially enabling OGSA-DAI

Throughout this dissertation, it has been mentioned that the ISO 19100 series of standards together with the OGC implementations have been implemented in a number of SDIs. To grid-enable these SDIs, would require grid-enabling these ISO standards and OGC implementation

specifications. Aloisio *et al.* (2005a) and Di *et al.* (2008) recently reported about an implementation for which OGC web services were grid-enabled. However, more of these implementations are required to better understand the challenges under different circumstances. Not only would such implementations increase the skills availability, they would also promote the development of tools to streamline these implementations, and ultimately provide input into standardizing grid-enabled components.

OGSA-DAI already provides uniform access to different relational databases, similar tot an OGC web service, which provides uniform access to different sources of geographic information. Future studies for OGSA-DAI could investigate uniform access for spatial data, with or without making use of OGC web services. Also, interesting would be a spatially enabled distributed query processing (DQP) of OGSA-DAI.

7.3.2 Developing an international address standard

In this dissertation an interoperable address data model, based on three principles, was presented in Chapter 4. The SANS 1883 street address type was described in terms of this data model. The model and its principles should be tested against other address standards, national as well as international, in order to refine the model so that it accommodates all standards. An interoperable address data model is an essential requirement for address data sharing. This requirement is already evident in the development of national address standards that have been successfully implemented for centralized collation of address data in countries such as the United Kingdom and Australia. To share and exchange address data on an international level, for example, as required in the disaster response scenario described in Chapter 1, an international address standard is required. Current international address standards are either too focused, such as the UPU-S42 for postal addresses, which does not cater for all purposes and types of addresses; or do not cater for address data as reference data, such as the one published by OASIS CIQ. A future study could investigate which option to follow: adapt existing international address standards to cater for all the above-mentioned requirements; develop an international address standard based on an existing standard such as 19112; or develop an international address standard from scratch.

7.3.3 Involving the community and trusting address data resources

The work in this dissertation is based on the assumption that the address data providers are mostly local authorities in an SDI that can be trusted to have address data of sufficient accuracy and quality. This assumption stems from the fact that address data providers are mostly local authorities that have a mandate to produce and maintain address data and are bound by regulations to produce this data according to certain agreed specifications and quality levels. However, in a Web 2.0 world, where the citizens become the sources for data, this assumption will not hold anymore. While

citizens, living at an address, are the best available source to verify an address, the question is whether they can be trusted to provide accurate data. Other researchers are also raising this question, such as Goodchild (2008) and Craglia *et al.* (2008). Future work could investigate how such a 'wikification' of address data can be securely and accurately integrated into Compartimos. The question is whether the elements of collective intelligence or crowd-sourcing that are present in these activities, in which contributors are able to challenge or edit the earlier contributions of others, is the modern equivalent of the process of consensus that the naming authorities have traditionally relied on and managed (Goodchild and Hill 2008).

7.3.4 Rolling out Compartimos in an SDI

Compartimos was designed for address data in an SDI. Further research could investigate how to extend Compartimos for other types of spatial data that is shared in an SDI. This work would have to center around the information viewpoint: how to integrate different datasets and make them interoperable, and how to extend the catalogue to cater for all kinds of information. Incorporating recent research findings on ontologies for interoperability would be relevant. A reference model for data grids that caters for all kinds of geographic information could be seen as the first step along the long path of standardizing geospatial data grids.

Compartimos focuses on the technical aspects of an SDI, i.e. the technologies, systems and standards. The non-technical aspects, such as policies, legislation, agreements, human and economic resources, and organizational aspects are beyond the scope of this work, but are a necessary next step in order to understand what it takes to grid-enable an SDI.

7.3.5 SDI in the clouds

The research on this dissertation was started in 2005, before the current hype of 'cloud computing'. However, clouds, such as those by Amazon, IBM, Microsoft and the like, also stand in line as the enabling platform for data sharing in an SDI. Instead of investing servers and bandwidth at different local authorities, local authorities could buy scalable computing power and data storage in a cloud. Apart from the on demand storage and processing capacity in the cloud, there is the further appeal that there is no need to support an IT infrastructure at the local authority. In a developing country such as South Africa, where shortages of IT skills are high, this approach would be worthwhile investigating. Thus, a future study could investigate the viability of data sharing in an SDI, that takes place in the clouds.

References

1. Aalders HJGL (2005). An introduction to spatial metadata standards in the world, in *World Spatial Metadata standards*, edited by Moellering H, Aalders HJGL and Crane A, Elsevier, Oxford, United Kingdom.
2. Acton D (2007). *Methods of charging for data in the NAD*, Hons project report, University of Pretoria, Pretoria, South Africa.
3. Arefin MA, Sadik MS, Coetzee S, Bishop JM (2006). Alchemi vs Globus: a performance comparison, *4th International Conference on Electrical and Computer Engineering*, December 19-21 2006, Dhaka, Bangladesh, pp602-605.
4. Aloisio G, Cafaro M, Conte D, Tiore S, Epicoco I, Marra GP, Quarta G (2005a). A Grid-Enabled Web Map Server, *International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I*, 2005, pp.298-303.
5. Aloisio G, Cafaro M, Fiore S, Wuarta G (2005b). A grid-based architecture for earth observation data access, *2005 ACM Symposium on Applied Computing*, March 13-17 2005, Santa Fe, New Mexico, USA.
6. Antonioletti M, Atkinson M, Baxter R, Borley A, *et al.* (2005). The design and implementation of Grid database services in OGSA-DAI, *Concurrency and Computation: Practice and Experience*, Vol. 17, Issue 2-4, pp.357-376.
7. Aydin G, Sayar A, Gadgil H, Aktas M, Fox GC, Ko S, Bulut H and Pierce ME (2008). Buliding and applying geographical information system Grids, *Concurrency and Computation: Practice and Experience*, **20**(14), pp 1653-1695.
8. Baker M, Apon A, Ferner C, Brown J (2005). Emerging grid standards, *IEEE Computer* April 2005, Vol. 38 No.4 pp.43-50.
9. Baranski B (2008). Grid Computing enabled WPS, paper accompanying the slides presented at the *GI Days* in Muenster, Germany, 16-18 June, 2008.
10. Belussi A, Brovelli MA, Negri M, Pelagatti G and Sanso F (2006). Dealing with multiple accuracy levels in spatial databases with continuous update, *Proceedings of Accuracy 2006 - 7th International Symposium on Spatial Accuracy Assesment in Natural Resources and Environmental Sciences*, Instituto Geografico Portugues, Lisboa, Portugal, pp 203-212.

11. Berman F, Fox G C, Hey A J G (ed.) (2003). *Grid Computing. Making the global infrastructure a reality*, John Wiley & Sons Ltd, Chichester, West Sussex PO19 8SQ, England.

12. Bernholdt D, Bharathi S, Brown D, *et al.* (2005). The Earth System Grid: Supporting the next generation climate modeling research, *Proceedings of the IEEE*, March 2005, **93**:5, pp485-495.

13. Bejar R, Latre A, Nogueras-Iso J, Muro-Medrano PR and Zaragoza-Soria FJ, (2008). An architectural style for spatial data infrastructures, *International Journal of Geographical Information Science*, 20 September 2008, available online ahead of print edition at http://www.tandf.co.uk/journals/tf/13658816.html, accessed 26 October 2008.

14. Buyya R and Nadiminti K (2006). Enterprise grid computing: State-of-the-art, *Enterprise Open Source Journal*, March/April 2006, pp19-22.

15. Brauner J and Schaeffer B (2008). Integration of GRASS functionality in web based SDI service chains, *Proceedings of the academic track of the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape Town, South Africa.

16. Brovelli MA, Magni D, Brioschi M, Legnani M and Corcoglioniti (2008). NAMGIS – A context-aware mobile Web GIS, *Proceedings of the academic Track of the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape Town, South Africa.

17. Bruin RP, White TOH, Walker AM, Austen KF, Dove MT, Tyer RP, Couch PA, Todorov IT and Blanchard MO (2008). Job submission to grid computing environments, *Concurrency and Computation: Practice and Experience*, **20**:1329-1340.

18. Carrera F and Ferreira J (2007). The Future of Spatial Data Infrastructures: Capacity-building for the Emergence of Municipal SDIs, *International Journal of Spatial Data Infrastructures Research (IJSDIR)*, 2007, Vol. 2, pp54-73.

19. Chervenak A, Foster I, Kesselman C, Salisbury C and Tuecke S (2000). The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets, *Journal of Network and Computer Applications,* July 2000, **23**(3), pp187–200.

20. Chervenak A, Schuler R, Kesselman C, Koranda S, Moe B (2005). Wide Area Data Replication for Scientific Collaborations, Grid Computing, 2005, *6th IEEE/ACM International Workshop on Grid Computing*, 13-14 Nov. 2005, pp1-8.

21. Chu X, Lonie A, Harris P, Thomas SR, Buyya R (2008). A service-oriented Grid environment for integration of distributed kidney models and resources, *Concurrency and Computation: Practice and Experience*, **20**:1095-1111.

22. Coetzee S (2008). Address data exchange in South Africa, *Proceedings of the ISO Workshop on address standards: Considering the issues related to an international address standard*, 25 May 2008, Copenhagen, Denmark.

23. Coetzee S and Bishop J (1998). A new way to query GIS on the web, *IEEE Software*, May/June 1998, **15**(3), pp31-40.

24. Coetzee S and Bishop J (2008). Address databases for national SDI: Comparing the novel data grid approach to data harvesting and federated databases, *International Journal of Geographic Information Science*, 26 September 2008, available online ahead of print edition at http://www.tandf.co.uk/journals/tf/13658816.html, accessed 26 October 2008.

25. Coetzee S and Cooper AK (2007a). The value of addresses to the economy, society and governance – a South African perspective, *45th Annual URISA Conference*, 20-23 August 2007, Washington DC, USA.

26. Coetzee S and Cooper AK (2007b). What is an address in South Africa? *South African Journal of Science*, Nov/Dec 2007, **103**(11/12), pp449-458.

27. Coetzee S and Cooper AK (2008). Can the South African address standard (SANS 1883) work for small local municipalities? *Proceedings of the academic Track of the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape Town, South Africa.

28. Coetzee S, Cooper AK, Lind M (ed.) (2008a). *Proceedings of the ISO Workshop on address standards – Considering the issues related to an international address standard*, 25 May 2008 Copenhagen, Denmark available online at http://www.isotc211.org/Address/Copenhagen_Address_Workshop/workshop.htm, accessed 19 June 2008.

29. Coetzee S, Cooper AK, Lind M, McCart Wells M, Yurman SW, Wells E, Griffiths N and Nicholson MJL (2008b). Towards an international address standard, *GSDI-10 Conference*, Trinidad and Tobago, 25 – 29 February 2008.

30. Colouris G, Dollimore J and Kindberg T (2005). *Distributed Systems – Concepts and Design*, Pearson Education, Essex, England, UK.

31. Cooper AK (1993). *Standards for exchanging digital geo-reference information*, MSc thesis, University of Pretoria, South Africa.

32. Cooper AK (2008). Overview of an address and purpose of the workshop, *Proceedings of the ISO Workshop on address standards: Considering the issues related to an international address standard*, 25 May 2008, Copenhagen, Denmark.

33. Cooper AK and Coetzee S (2008). The South African address standard and initiatives towards an international address standard, *Proceedings of the academic track of the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape Town, South Africa.

34. Cohen J, Darlington J and Lee W (2008), Payment and negotiation for the next generation Grid and Web, *Concurrency and Computation: Practice and Experience*, 2008, **20**:239-251.

35. Coppola M, Jégou Y, Morin C, Prieto LP, Sánchez OD, Yang EY and Yu H (2008). Virtual organization support within a grid-wide operating system, *IEEE Internet Computing*, March/April 2008, vo.12, no. 2, pp20-28.

36. Craglia M, Goodchild MF, Annoni A, Camara G, Gould M, Kuhn W, Mark D, Masser I, Maguire D, Liang S and Parsons E (2008), Next-Generation Digital Earth, a position paper from the Vespucci Initiative for the advancement of Geographic Information Science, *International Journal of Spatial Data Infrastructures Research*, 2008,Volume 3, pp146-167.

37. Crompvoets J, Bregt A, Rajabifard A, Williamson I. (2004). Assessing the worldwide developments of national spatial data clearinghouses, *International Journal of Geographical Information Science*, October-November 2004, vol. 18, no. 7, pp665-689.

38. Davis CA Jr and Fonseca T (2007). Assessing the certainty of locations produced by an address geocoding system, *Geoinformatica*, **11**:103-129.

39. De Bree F, Eertink D, Laarakker P (2008). Assessing Quality of Collaboration in Netherlands SDI, *GSDI-10 Conference*, 25-29 February 2008, St Augustine, Trinidad.

40. Delic KA and Walker MA (2008), Emergence of the academic computing clouds, *ACM Ubiquity*, **9**(31), 5-11 August 2008, no page numbers available.

41. De Man WHE (2006). Understanding SDI; complexity and institutionalization, *International Journal of Geographical Information Science*, March 2006, vol. 20, no. 3, pp329-343.

42. De Man WHE (2007). Beyond Spatial Data Infrastructures there are no SDIs – so what, *International Journal of Spatial Data Infrastructures Research (IJSDIR)*, 2007, vol. 2, pp1-23.

43. De Rose CAF, Ferreto T, Calheiros RN, Cirne W, Costa LB and Fireman D (2008). Allocation strategies for utilization of space-shared resources in Bag of Tasks grids, *Future Generation Computer Systems*, May 2008, vol. 24, no. 5, pp331-341.

44. Di L, Chen A, Yang W, Zhao P (2003). The integration of Grid technology with OGC Web services (OWS) in NWGISS for NASA EOS Data, *GGF8 & HPDC12*, 24 – 27 June 2003, at Seattle, SE , USA .

45. Di L, Chen A, Yang W, Liu Y, Wei Y, Mehrotra P, Hu C, Williams D (2008). The development

of a geospatial data Grid by integrating OGC web services with Globus-based Grid technology, *Concurrency and Computation: Practice and Experience*, September 2008, **20**(14), pp1617-1635.

46. Farvacque-Vitkovic C, Godin L, Leroux H, Chavez R and Verdet F (2005). *Street Addressing and the Management of Cities*, Washington DC, US: The World Bank.
47. Finnigan JV and Blanchette J (2008). A Forward-Looking Software Reuse Strategy, *2008 IEEE Aerospace Conference*, 1-8 March 2008, pp1 – 9.
48. Foster I (2002). What is the Grid? A three point checklist, *GRIDToday*, Vol. 1 No. 6, July 22 2002.
49. Foster I (2003). The Grid: A new infrastructure for the 21st century science in *Grid Computing. Making the global infrastructure a reality*, by Berman F, Fox G C, Hey A J G (ed.), 2003, John Wiley & Sons Ltd, Chichester, West Sussex PO19 8SQ, England.
50. Foster I and Kesselman C (1999). Preface in *The GRID: Blueprint for a New Computing Infrastructure*, edited by Ian Foster and Carl Kesselman, Morgan Kaufmann Publishers Inc., San Francisco, California, USA.
51. Foster I and Kesselman C (2004). Concepts and Architecture, in *The GRID 2: Blueprint for a New Computing Infrastructure*, edited by Ian Foster and Carl Kesselman, Morgan Kaufmann Publishers Inc, San Francisco, California, USA.
52. Foster I, Kesselman D, Nick JM and Tuecke S (2002). *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*, available online at http://forge.gridforum.org/sf/go/doc13483?nav=1, accessed 8 November 2008.
53. Foster I, Kesselman C and Tuecke S (2001). The Anatomy of the Grid – enabling scalable virtual organizations, *International Journal of High Performance Computing Applications*, 15(3), pp200-222.
54. Foster I, Williams DN, Middleton D (2006). *Earth System Grid II - Turning Climate Datasets into Community Resources*, Final report for the period June 1, 2001 – July 31, 2006, available online at http://datagrid.ucar.edu/esg/about/docs/ESG_II_Final_Report.doc, accessed 8 November 2008.
55. Georgiadou SK, Puri SK and Sahay S (2005). Towards a potential research agenda to guide the implementation of spatial data infrastructures – A case study from India, *International Journal of Geographical Information Science*, November 2005, Vol. 19, No. 10, pp.1113-1130.
56. Ghimire D, Simonis I and Wytszisk A (2005). Integration of grid approaches into the geographic Web services domain, *FIG Working Week 2005 and GSDI-8*, Cairo, Egypt, April

16-21 2005.

57. Gomez-Iglesias A, Vega-Rodriguez MA, Castejon-Magana F, del Solar MR and Montes MC (2008). Grid Computing in Order to Implement a Three-Dimensional Magnetohydrodynamic Equilibrium Solver for Plasma Confinement, *16th Euromicro Conference on Parallel, Distributed and Network-Based Processing (PDP 2008)*, 13-15 February 2008, pp435 – 439.
58. Goodchild MF (2008). Spatial accuracy 2.0, in J.-X. Zhang and M.F. Goodchild, editors, *Spatial Uncertainty*, Proceedings of the Eighth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Volume 1. Liverpool: World Academic Union, pp1–7
59. Goodchild MF and Hill L (2008). Introduction to digital gazetteer research, International Journal of Geographic Information Science, **22**(10), pp 1039-1044.
60. Goodchild MF, Johnston DM, Maguire DJ, Noronha VT (2005). Distributed and mobile computing, in *A research agenda for geographic information science*, edited by RB McMaster and EL Usery, CRC Press, Boca Raton, Florida, USA.
61. Granell C, Díaz L, Gould M (2007), Managing earth observation data with distributed geoprocessing services, *IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2007)*, 23-28 July 2007, pp4777 – 4780.
62. Grimshaw D and Natrajan A (2005). Legion: Lessons Learned Building a Grid Operating System, *Proceedings of the IEEE*, March 2005, vol. 93, no. 3, pp589-603.
63. Harvey F. and Tulloch D (2006). Local-government data sharing: Evaluating the foundations of spatial data infrastructures, *International Journal of Geographical Information Science*, August 2006, **20**(7), pp743-768.
64. Hayes B (2008). Cloud computing, *Communications of the ACM*, July 2008, **51**(7), pp9-11
65. Hong S-K (2008). Ubiquitous geographic information (UBGI) and address standards, *Proceedings of the ISO Workshop on address standards – Considering the issues related to an international address standard*, 25 May 2008, Copenhagen, Denmark.
66. Hua L, De-ren L, Xin-yan Z (2005). Large volume spatial data management based on grid computing, *Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium (IGARSS '05)*, 25-29 July 2005.
67. Iglesias CA (2008). SDI in Chile – National System of Coordination of Territorial Information (SNIT) – State of the art and projections, *GSDI-10 Conference*, Trinidad and Tobago, 25 – 29 February 2008.
68. Jacoby S, Smith J, Ting L, and Williamson I (2002). Developing a common spatial data

infrastructure between state and local government—an Australian case study, *International Journal of Geographical Information Science*, June 2002, Vol. 6 No 4, pp. 305-322.

69. Jang S-G and Kim TJ (2006). Modeling an interoperable multimodal travel guide system using the ISO 19110 series of international standards, *Proceedings of the ACM GIS 2006 Conference*, 10-11 November 2006, Arlington, Virginia, USA.
70. Jin-Hsiang S and Chung-Chi C (2008). The Development of Taiwan Geospatial One-Stop Portal, *GSDI-10 Conference*, Trinidad and Tobago, 25 – 29 February 2008.
71. Lanig A and Zipf A (2008). Requirements for efficient mining and processing of massive terrain data in Grid infrastructures, *Geophysical Research Abstracts*, EGU General Assembly 2008, Vol. 10.
72. Lawton G (2008). Moving the OS to the Web, *IEEE Computer*, March 2008, vol. 41, no. 3, pp. 16-19.
73. Liang Z, Tang X and Lan W (2007). 'Digital China' Geo-spatial framework construction: situation, problems and suggestions, *International Conference on Wireless Communications, Networking and Mobile Computing*, 21-25 September 2007, pp4995 – 4998.
74. Li H and Buyya R (2007). Model-Driven Simulation of Grid Scheduling Strategies. *Proceedings of the IEEE International Conference on e-Science and Grid Computing*, Bangalore India, 10-13 Dec. 2007, pp287-294.
75. Masser I, Rajabifard A, Williamson I (2007). Spatially enabling governments through SDI implementation, *International Journal of Geographic Information Science*, January 2008, vol. 22, no.1, pp 5-20.
76. Matheri M. (2005). Challenges facing the creation of a standard South African address system, *FIG Working Week and 8th Global Spatial Data Infrastructure Conference (GSDI-8)*, 16-21 April 2005, Cairo, Egypt.
77. McDougall K, Rajabifard A and Williamson I (2005). What will motivate local governments to share spatial information?, *SSC 2005 Spatial Intelligence, Innovation and Praxis: The national biennial Conference of the Spatial Sciences Institute*, September 2005, Melbourne, Australia.
78. Molina M and Bayarri S (2008). The Andean Information System for Disaster Prevention and Relief: a case study of multi-national open-source SDI, *Proceedings of the academic track of the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape Town, South Africa.
79. Morad M (2002). British standard 7666 as a framework for geocoding land and property information the UK, *Computers, Environment and Urban Systems*, September 2002, vol. 26

no.5, pp483-492.

80. Morad S and Kuflik T (2005). Conventional and open source software reuse at Orbotech - an industrial experience, *Proceedings of the IEEE International Conference on Software - Science, Technology and Engineering*, 22-23 February 2005, pp110 – 117.

81. Olivier MS (1999). *Information Technology Research. A Practical Guide*, MS Olivier, 1999.

82. Onsrud H, Poore B, Rugg R, Taupier R, Wiggins L (2005). The future of the spatial information infrastructure, in *A research agenda for geographic information science*, edited by RB McMaster and EL Usery, CRC Press, Boca Raton, Florida, USA.

83. Peng Z-R and Tsou M-H (2003). *Internet GIS – Distributed geographic information services for the Internet and wireless networks*, John Wiley & Sons, Hoboken, New Jersey, USA.

84. Plaszczak P and Wellner R Jr (2006). *Grid Computing – The Savvy Manager's Guide*, Morgan Kaufmann Publishers Inc, San Francisco, California, USA.

85. Rabl T, Pfeffer M and Kosch H (2008). Dynamic allocation in a self-scaling cluster database, *Concurrency and Computation - Practice and Experience,* **20**(17), pp2025-2038.

86. Rahed AA, Coetzee S and Rademeyer M (2008). A data model for efficient address data representation - Lessons learnt from the Intiendo address matching tool, *Proceedings of the academic track of the 2008 Free and Open Source Software for Geospatial (FOSS4G) Conference, incorporating the GISSA 2008 Conference*, 29 September - 3 October 2008, Cape Town, South Africa.

87. Rajabifard A, Feeney MF, Williamson I and Masser I (2003), *National spatial data infrastructures.* In Developing Spatial Data Infrastructures: from Concept to Reality, I Williamson, A Rajabifard and M-E.F. Feeney (Eds), pp. 95–109, Taylor & Francis, London, UK.

88. Rajabifard A (2008). Re-engineering SDI design to support spatially enabled society, *INSPIRE Conference*, 23 June 2008, Maribor, Slovenia.

89. Rajabifard A, Binns A and Williamson I (2005). Creating an enabling platform for the delivery of spatial information. *Proceedings of SSC 2005 Spatial Intelligence, Innovation and Praxis: The national biennial Conference of the Spatial Sciences Institute*, Melbourne, Australia, September 2005, Melbourne: Spatial Sciences Institute, ISBN 0-9581366-2-9.

90. Rajabifard A, Binns A, Masser I and Williamson I (2006). The role of sub-national government and the private sector in future spatial data infrastructures, *International Journal of Geographical Information Science*, August 2006, Vol.20, No.7, pp.727-741.

91. Ripeanu M, Singh MP, Vazhkudai SS (2008), Virtual Organizations, *IEEE Internet Computing*,

March/April 2008, **12**(2), pp10-12.

92. Schaeffer B and Baranski B (2008). Orchestrating grid computing enabled web processing services, *2008 Geoinformatics Conference,* Potsdam, Germany, 11-13 June 2008, abstract available online at http://gsa.confex.com/gsa/2008GE/finalprogram/abstract_142272.htm, accessed 8 September 2008.
93. Sheth AP and Larson JA (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases, *ACM Computing Surveys*, September 1990, **22**(3), pp183-236.
94. Shu Y, Zhang JF, Zhou X (2006). A Grid-Enabled Architecture for Geospatial Data Sharing, *IEEE Asia-Pacific Conference on Services Computing* (APSCC '06), December 2006, pp369 – 375.
95. Singh G, Bharathi, Chervenak A, Deelman E, Kesselman C, Manohar M, Patil S and Pearlman L (2003). A Metadata Catalog Service for Data Intensive Applications, *Proceedings of the 2003 ACM/IEEE conference on Supercomputing (SC '03)*, November 2003.
96. Talia D (2002). The Open Grid Services Architecture: Where the Grid Meets the Web, *IEEE Internet Computing*, November/December 2002, **6**(6), pp 67-71.
97. Tracz W (1994). Software reuse myths revisited, Software Engineering, 1994. *Proceedings of the 16th International Conference on Software Engineering (ICSE-16)*, 16-21 May 1994, pp271 – 272.
98. Tuladhar A, Radwan M, Kader F and El-Ruby S (2005). Federated data model to improve accessibility of distributed cadastral databases in land administration, *Proceedings of 8th Global Spatial Data Infrastructure Conference (GSDI-8)*, 16-21 April 2005, Cairo, Egypt.
99. Valentin A and Cabello M (2008). SDI's and Knowledge, *GSDI-10 Conference*, Trinidad and Tobago, 25 – 29 February 2008.
100. Venugopal S, Buyya R and Ramamohanarao K (2006). A taxonomy of data grids for distributed data sharing, management and processing, *ACM Computing Surveys*, March 2006, Vol. 38, Article 3, pp.1-53.
101. Volckaert B, Wauters T, De Leenheer M, Thysebaert P, De Turck F, Dhoedt B and Demeester P (2008), Gridification of collaborative audiovisual organizations through the MediaGrid framework, *Future Generation Computer Systems*, vol. 24, no. 5, pp371-389.
102. Wang Y, GE L, Rizos C, & Babu R (2004). Spatial data sharing on grid, *Geomatics Research Australasia*, 81, 3-18.
103. Wei X, Yue P, Dadi U, Min M, Hu C, Di L (2006). Effective Acquisition of Geospatial Data

Products in a Collaborative Grid Environment, *IEEE International Conference on Services Computing*, SCC '06, Sept. 2006, pp455-462.

104. Weiss A (2007). Computing in the clouds, *netWorker*, December 2007, **11**(4), pp16-25.

105. Wells M, Anderson C, Perkins H, Wells E & Yurman S (2008). Developing a Comprehensive Standard for US Address Data, *Proceedings of the ISO Workshop on address standards: Considering the issues related to an international address standard*, 25 May 2008, Copenhagen, Denmark.

106. Williamson I, Grant D and Rajabifard A (2005). Land administration and spatial data infrastructures, *Proceedings of 8th Global Spatial Data Infrastructure Conference (GSDI-8)*, 16-21 April 2005, Cairo, Egypt.

107. Williamson I, Rajabifard A and Binns A (2006). Challenges and issues for SDI development, *International Journal of Spatial Data Infrastructures Research (IJSDIR)*, 2006, vol. 1, pp24-35.

108. Wladawsky-Berger I (1998). "The Industrial Imperative" in *The GRID Blueprint for a new computing infrastructure*, pp25-34, Morgan Kaufmann Publishers Inc., San Francisco, California, USA.

109. Wytzisk A, Von Dömming A, Voges U (2008). Technical Architecture and Implementation Plan for GDI-DE, *GSDI-10 Conference*, Trinidad and Tobago, 25 – 29 February 2008.

110. Xue Y, Wan W, Li Y, Guang J, Bai L, Wang Y, and Ai J (2008). Quantitative retrieval of geophysical parameters using satellite data, *IEEE Computer*, April 2008, **41**(4), pp33-39.

111. Yildirim V AND Yomralioglu T (2004). An address-based geospatial application. *FIG Working Week*, 22-27 May 2004, Athens, Greece.

112. Zaslavsky I, He H, Tran J, Martone M E, Gupta A (2004). Integrating brain data spatially: spatial data infrastructure and atlas environment for online federation and analysis of brain images, *Proceedings on the 15th International Workshop on Database and Expert Systems Applications*, 30 August – 3 September 2004, pp389-393.

113. Zhao P, Chen A, Liu Y, Di L, Yang Q, Li P (2004). Grid Metadata Catalog Service-Based OGC Web Registry Service, *GIS'04*, November 12–13, 2004, Washington D. C., USA.

114. Zhao P, Yu G and Di L (2007). Geospatial Web services, in *Emerging spatial information systems and applications*, edited by Brian H. Hilton, Idea Group Publishing, Hershey, PA, USA.

Referenced Standards

1. *Draft Street Address Standard* (2005), US Address Standard Working Group, United States Federal Geographic Data Committee, US.
2. AS/NZS:4819:2003 (2003). *Geographic information – rural and urban addressing*, jointly published by Standards Australia, Sydney, Australia, and Standards New Zealand, Wellington, New Zealand.
3. BS 7666-0:2006 (2006). *Spatial datasets for geographical referencing - Part 0: General model for gazetteers and spatial referencing*, British Standards Institute (BSI), London, UK.
4. ISO 11180:1993 (1993). *Postal addressing* (withdrawn), International Organization for Standardization (ISO), Geneva, Switzerland.
5. ISO 15836:2003 (2003). *Information and documentation - The Dublin Core metadata element set*, International Organization for Standardization (ISO), Geneva, Switzerland.
6. ISO 19101:2002 (2002). *Geographic information – Reference model*, International Organization for Standardization (ISO), Geneva, Switzerland.
7. ISO 19111:2007 (2007). *Geographic information – Spatial referencing by coordinates*, International Organization for Standardization (ISO), Geneva, Switzerland.
8. ISO 19112:2003 (2003). *Geographic information – Spatial referencing by geographic identifiers*, International Organization for Standardization (ISO), Geneva, Switzerland.
9. ISO 19115:2003 (2003). *Geographic information - Metadata*, International Organization for Standardization (ISO), Geneva, Switzerland.
10. ISO 19116:2004 (2004). *Geographic information - Positioning services*, International Organization for Standardization (ISO), Geneva, Switzerland.
11. ISO 19117:2005 (2005). *Geographic information - Portrayal*, International Organization for Standardization (ISO), Geneva, Switzerland.
12. ISO 19119: 2005 (2005). *Geographic information –Services*, International Organization for Standardization (ISO), Geneva, Switzerland.
13. ISO 19136:2007 (2007). *Geographic information – Geography Markup Language (GML)*, International Organization for Standardization (ISO), Geneva, Switzerland.
14. ISO 19142 (draft) (2008). *Geographic information – Web Feature Service*, International

Organization for Standardization (ISO), Geneva, Switzerland.

15. ISO 21127:2006 (2006). *Information and documentation - A reference ontology for the interchange of cultural heritage information*, International Organization for Standardization (ISO), Geneva, Switzerland.
16. ISO 3166-1:2006 (2006). *Codes for the representation of names of countries and their subdivisions – Part 1: Country codes*, International Organization for Standardization (ISO), Geneva, Switzerland.
17. ISO/IEC 10746:1998 (1998), *Information technology - Open Distributed Processing - Reference Model,* International Organization for Standardization (ISO), Geneva, Switzerland.
18. ISO/IEC 18026:2006 (2006), *Information technology – Spatial Reference Model (SRM),* International Organization for Standardization (ISO), Geneva, Switzerland.
19. ISO/IEC 7498 (1994), *Information technology - Open Systems Interconnection – Basic Reference Model,* International Organization for Standardization (ISO), Geneva, Switzerland.
20. OASIS Customer Information Quality (QIC) TC (2007). *OASIS CIQ V3.0 Committee DRAFT Specifications*, available online at http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=ciq#download, accessed 8 November 2008.
21. Open Geospatial Consortium (2005). *Abstract Specifications*, available online at http://www.opengeospatial.org/standards/as, accessed 28 March 2008.
22. Open Geospatial Consortium (2007). *OpenGIS Catalogue Service Implementation Specification 2.0.2*, available online at http://www.opengeospatial.org/standards/cat, accessed 10 November 2008.
23. Open Geospatial Consortium (2003). *OGC Reference Model*, available online at http://www.opengeospatial.org/standards, accessed 17 April 2008.
24. Open Geospatial Consortium (2005). *OpenGIS Web Feature Service Implementation Specification 1.1.0*, available online at http://www.opengeospatial.org/standards/wfs, accessed 28 March 2008.
25. Open Geospatial Consortium (2003). *OpenGIS Web Map Service Implementation Specification 1.3.0*, available online at http://www.opengeospatial.org/standards/wms, accessed 28 March 2008.
26. Open Geospatial Consortium (2007). *OpenGIS Web Processing Service Implementation Specification 1.0.0*, available online at http://www.opengeospatial.org/standards/wps, accessed 28 March 2008.

27. SANS/WD 1883-1 (2008). *Geographic information – Address Standard, Part 1: Data format of addresses (committee draft)*, Standards South Africa, Pretoria.

28. SANS/WD 1883-2 (2008). *Geographic information – Address Standard, Part 2: Guidelines for addresses in databases, data transfer, exchange and interoperability (committee draft)*, Standards South Africa, Pretoria.

29. SANS/WD 1883-3 (2008). *Geographic information – Address Standard, Part 3: Guidelines for address allocation and updates (committee draft)*, Standards South Africa, Pretoria.

30. UPU S42 (2006). *S42: International postal address components and templates*, Universal Postal Union (UPU), Berne, Switzerland.

Other references

Below is a list of other documents, including but not limited to reports, acts, directives, projects and websites, that are referenced in this dissertation.

1. *52°North web site*, available online at http://52north.org/, accessed 9 September 2008.
2. AfriGIS (2008). *Data Release Notes*, AfriGIS, 30 August 2008.
3. AfriGIS Website, available online at www.afrigis.co.za, accessed 4 November 2008.
4. Amazon (2008). *Announcing Elastic IP Addresses and Availability Zones for Amazon EC2, Amazon Web Services - What's new?,* available online at http://aws.amazon.com/about-aws/whats-new/, accessed 4 November 2008.
5. Brady M, Gavaghan D, Harris S, Jirotka M, Knox A, Lloyd S *(*2005). *E-DiaMoND: The Blueprint Document*, available online on the eDiaMoND grid computing project website at http://www.ediamond.ox.ac.uk/publications/blueprint-Final.pdf, accessed 8 November 2008.
6. Cambridge University Press *(*2007). *Online edition of the Cambridge Advanced Learner's Dictionary*, available online at http://dictionary.cambridge.org/define.asp?key=982&dict=CALD, accessed 8 November 2008.
7. Chief Directorate: Surveys and Mapping website (2004)*, WGS84* page, available online at http://w3sli.wcape.gov.za/SURVEYS/MAPPING/wgs84.htm, accessed 3 November 2008.
8. *Constitution of the Republic of South Africa (*1996). Available online at http://www.polity.org.za/polity/govdocs/constitution/index.html, accessed 19 March 2008.
9. *Dictionary.com* (2008). Published by Lexico Publishing Group, LLC, available online at http://dictionary.reference.com/, accessed 5 November 2008.
10. *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community* (INSPIRE) [2007] OJ l 108/1, available online at http://inspire.jrc.it/directive/l_10820070425en00010014.pdf, accessed 28 March 2008.
11. *Dublin Core Metadata Initiative Website*, The Dublin Core Metadata Initiative, available online at www.dublincore.org, accessed 10 November 2008.
12. Fahey D and Finch F *(*2006). *GeoDirectory Technical Guide*. An Post GeoDirectory Limited, available online at http://www.geodirectory.ie/Downloads.aspx, accessed 19 March 2008.

13. *Financial Intelligence Centre Act of South Africa (*2001). Available online at http://www.acts.co.za/fica/index.htm, accessed 19 March 2008.

14. *Geodateninfrastruktur-Grid (GDI-Grid) project*, available online at http://www.d-grid.de/index.php?id=398&L=1, accessed 28 March 2008.

15. *GeoDirectory Website*. An Post GeoDirectory Limited, available online at http://www.geodirectory.ie, accessed 19 March 2008.

16. *Geosciences Network (GEON) portal*, available online at www.geongrid.org, accessed 30 January 2008.

17. Geosciences Network (GEON) project *(*2005). *GEON – Cyberinfrastructure for the Geosciences: 2005 Annual Report,* available online at http://www.geongrid.org/communications/annual_reports/Annual_Report_2005_Final_NSF.pdf , accessed 30 January 2008.

18. Google (2008). *Google Site Search Taps Power of the Cloud to Improve Search for Business Websites*, Press Release available online at http://www.google.com/intl/en/press/pressrel/20080603_site_search.html, accessed 4 November 2008.

19. Haas H and Brown A (Ed.) (2004), *Web Services Glossary*. W3C Working Group Note 11 February 204, available online at http://www.w3.org/TR/ws-gloss/, accessed 21 October 2008.

20. *Knowledge Factory Website*, available online at http://www.knowledgefactory.co.za, accessed 3 November 2008.

21. IBM 2008, *IBM Launches Cloud Services Initiative*, Press Release available online at http://www-03.ibm.com/press/us/en/pressrelease/25341.wss, accessed 4 November 2008.

22. *International Organization for Standardization (ISO)* Press Release (2008), available online at http://www.iso.org/iso/pressrelease.htm?refid=Ref1110, accessed 28 March 2008.

23. *International Organization for Standardization Technical Committee 211 (ISO/TC 211) Website*, available online at www.isotc211.org, accessed on 28 March 2008.

24. International Organization for Standardization Technical Committee 211 (ISO/TC 211) *(*2003). *Terms of Reference Joint Advisory Group between ISO/TC 211 and OGC*, Document no. 1513, available online at http://www.isotc211.org/protdoc/211n1531/, accessed 8 November 2008.

25. *Laser Interferometer Gravitational Wave Observatory (LIGO)* website, available online at http://www.ligo.caltech.edu/, accessed 8 November 2008.

26. Levoleger K and Corbin C *(*Ed.) *(*2005). *Survey of European National Addressing as of May*

2005, European Umbrella Organisation for Geographic Information (EUROGI), available online at http://eurog.vbnprep.com/POOLED/DOCUMENTS/a101730/EUROGI_Address_Survey_Resp_V3Final.pdf, accessed 28 March 2008.

27. Lind M *(*2004). Reliable address data: Developing a common address reference system, in Section 6: Reference Data of *A Compendium of SDI Best Practice* edited by C Corbin, available online at http://www.ec-gis.org/ginie/documents.html, accessed 8 November 2008.
28. Lind M and Nicholson MJL *(*2004). A database of reference for national addressing. Denmark and England and Wales compares, in Section 6: Reference Data of *A Compendium of SDI Best Practice* edited by C Corbin, available online at http://www.ec-gis.org/ginie/documents.html, accessed 8 November 2009.
29. Mathur A (2008). Email, 27 October 2008.
30. Microsoft (2008). *Microsoft Unveils Windows Azure at Professional Developers Conference*, Press Release available online at http://www.microsoft.com/presspass/press/2008/oct08/10-27PDCDay1PR.mspx, accessed 4 November 2008.
31. National Survey and Cadastre and Danish Agency for Enterprise and Construction *(*2005). *Offentlige adresse-webservices* (en: Joint Address Web Services), available online at http://www.adresse-info.dk/Portals/2/Dok/Adresse-webservices_Business_case_undersoegelse.pdf, accessed 8 November 2008.
32. OASIS (2008). *OASIS SOA Reference Model (SOA-RM) FAQ*, available online at http://www.oasis-open.org/committees/soa-rm/faq.php, accessed 5 November 2008.
33. OGF Transaction Management Research Group, (2005). *Proposed Grid Transactions RG – Charter*, Open Grid Forum, available online at http://www.ogf.org/tm-rg-charter.html, accessed September 2006.
34. *OGSA-DAI Website*. University of Edinburgh, available online at http://www.ogsadai.org.uk/, accessed 19 March 2008.
35. Open Geospatial Consortium (OGC) *(*2008a). *OGC® and OASIS Announce Progress on Standards Cooperation*, Press Release available online at http://www.opengeospatial.org/pressroom/pressreleases/849, accessed 28 March 2008.
36. Open Geospatial Consortium (OGC) (2008b). *Summary of the OGC Web Services, Phase 5 (OWS-5) Interoperability Testbed*, Editors: J Cook and R Singh, document reference number: OGC-08-073r2 Version 0.9.0, OGC, 22 July 2008, available online at http://portal.opengeospatial.org/files/?artifact_id=27995, accessed 3 November 2008.
37. Open Geospatial Consortium (OGC) (2008), *OGC Newsletter*, July 2008, available online at

http://www.opengeospatial.org/pressroom/newsletters/200807, accessed 4 November 2008.

38. *Open Geospatial Consortium (OGC) Website*, available online at www.opengeospatial.org, accessed 28 March 2008.

39. *Open Grid Forum (OGF) Website*, available online at www.opengridforum.org, accessed 10 September 2008.

40. Open Grid Forum (OGF) (2008a). *Defining the Grid: A Roadmap for OGSA Standards, Version 1.1*, Editors: C. Jordan and H. Kishimoto, available online at http://www.ogf.org/gf/docs/?final, accessed 2 April 2008.

41. Open Grid Forum (OGF) (2008b). *OGC-OGF Collaboration Workshop Report*, OGF-22, 26 February 2008, available online at www.ogf.org/OGF22/materials/1075/OGC-OGF-Workshop-Report_OGF-22.doc, accessed 4 November 2008.

42. Open Grid Forum (OGF) (2007a). *OGSA Data Architecture*, Editors: D Berry, A Luniewski, M Antonioletti, available online at http://www.ogf.org/gf/docs/?final, accessed 2 April 2008.

43. Open Grid Forum (OGF) (2007b). *OGSA Data Architecture Scenarios, Version 0.16*, Editors: S Davey, A Anjomshoaa, M Antonioletti, M Atkinson, D Berry, A Chervenak *et al.* available online at http://www.ogf.org/gf/docs/?final, accessed 2 April 2008.

44. Open Grid Forum (OGF) (2007c). *Open Grid Services Architecture Glossary of Terms Version 1.6*, Editor: J. Treadwell, available online at http://www.ogf.org/gf/docs/?final, accessed 2 April 2008.

45. Open Grid Forum (OGF) (2006). *The Open Grid Services Architecture, Version 1.5*, Editors: I. Foster, H. Kishimoto and A. Savva, available online at http://www.ogf.org/gf/docs/?final, accessed 2 April 2008.

46. Oxford University Press (2007a). *Online edition of the Compact Oxford English Dictionary*, available online at http://www.askoxford.com/concise_oed/address?view=uk, accessed 18 November 2007.

47. Oxford University Press (2007b). *Online edition of the Oxford English Dictionary*, available online at http://dictionary.oed.com/cgi/entry/50002492?query_type=word&queryword=address&first=1&max_to_show=10&sort_type=alpha&search_id=qwq2-eYh0se-10712&result_place=1, accessed 8 November 2008.

48. Paull D (2003). *A Geocoded National Address File for Australia: The G-NAF What, Why, Who and When*. PSMA Australia Limited, available online at http://www.psma.com.au/resources/the-g-naf-what-why-who-and-when , accessed 19 March 2008.

49. *Public Sector Mapping Agencies (PSMA) Australia Website*, available online at www.psma.com.au, accessed 19 March 2008.

50. Rase D, Björnsson A, Probert M, Haupt M *(*Ed.) *(*2002). *INSPIRE Reference Data and Metadata Position Paper*. INSPIRE Reference Data and Metadata working group, available online at http://www.ec-gis.org/inspire/reports/position_papers/inspire_rdm_pp_v4_3_en.pdf, accessed 19 March 2008.

51. *RM-ODP: The Reference Model for Open Distributed Processing*, The RM-ODP Resources Site of ISO, IEC and ITU, available online at http://www.rm-odp.net/, last accessed 4 April 2008.

52. *Spatial Data Infrastructure Act of South Africa (*2003). Available online at http://llnw.creamermedia.co.za/articles/attachments/01109_spatdatinfraa54.pdf, accessed 19 March 2008.

53. *US Executive Order 12906* (1994). Published 13 April, 1994, edition of the Federal Register, Volume 59, Number 71, pp. 17671-17674.

Appendix A. Acronyms and abbreviations

Table 39. List of abbreviations used in this dissertation

ANZLIC	Spatial Information Council of Australia and New Zealand
ArcIMS	Arc Internet Map Server, a web map server product by the ESRI company
ArcView	Entry level GIS desktop software by the ESRI company
BIRN	Biomedical information research network
BCU	eDiaMoND breast care centers
CAD	Computer aided design
CEN	European Committee for Standardization
DBMS	Database management system
DOE	US Department of Energy
EBF	Extended Backus Naur Form
ERC	Emergency Response Center
ESG	Earth System Grid
ESRI	Environmental Systems Research Institute
FGDC	Federal geographic data committee
GEON	Geoscience network
GIS	Geographic information system
GISc	Geographic Information Science
GML	Geography markup language
G-NAF	Geocoded national address file of Australia
GPS	Global positioning system
GridFTP	An extension of the standard file transfer protocol (FTP) for use with grid

	computing
GUI	Graphical user interface
HTTP	Hypertext transfer protocol
IEC	International Electro-technical Commission
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
INSPIRE	**IN**frastructure for **SP**atial **I**nfo**R**mation in the **E**uropean Community
ISO	International Organization for Standardization
ISO/TC 211	ISO Technical Committee 211, Geographic information/Geomatics
ISP	Internet Service Provider
ITU-T	International Telecommunication Union's Telecommunication Standardization
JDBC	Java database connectivity
LIGO	Laser Interferometer Gravitational Wave Observatory
LIDAR	Light Detection and Ranging
MoU	Memorandum of Understanding
NASA	National Aeronautics and Space Administration
NLPG	National Land and Property Gazetteer
PSMA	Public Sector Mapping Agencies
RM-ODP	ISO Reference Model for Open Distributed Processing
SABS	South African Bureau of Standards
SAPO	South Africa Post Office
SDI	Spatial data infrastructure
SHP	ESRI Shapefile, or simply a shapefile, a popular geospatial vector data format for geographic information systems software
OASIS	Organization for the Advancement of Structured Information Standards
ODBC	Open database connectivity
OGC	Open Geospatial Consortium

OGF	Open Grid Forum
OGSA	Open Grid Services Architecture
OGSA-DAI	OGSA Data access and integration
OGSA-DQP	OGSA-DAI distributed query processing
OGSA-GDS	OGSA-DAI grid data services
PKI	Public key infrastructure
PoP	Point of presence
RAID	Redundant array of indexed drives, or Redundant array of independent disks
SAPO	South African Post Office
SOA	Service-oriented architecture
SOAP	A protocol for exchanging XML-based messages over computer networks which once stood for simple object access protocol but this acronym was dropped with Version 1.2 of the standard, as it was considered to be misleading
TCP/IP	Transmission Control Protocol and Internet Protocol, the set of communications protocols that implement the protocol stack on which the Internet and most commercial networks run
THRIP	Technology and human resources for industry project
UDDI	Universal description discovery and integration
UK	United Kingdom
UML	Unified modeling language
UPU	Universal Postal Union
USA	United States of America
VO	Virtual organization
W3C	World Wide Web Consortium
WS	Web services
WCS	Web Catalogue Service
WFS	Web Feature Service
WGS84	World Geodetic System ellipsoid for 1984

WMS	Web Map Service
WPS	Web Processing Service
WSDL	Web services description language
WSRF	Web services resource framework
XML	Extensible Markup Language

Appendix B. Compartimos data

B.1 Data model: Address data

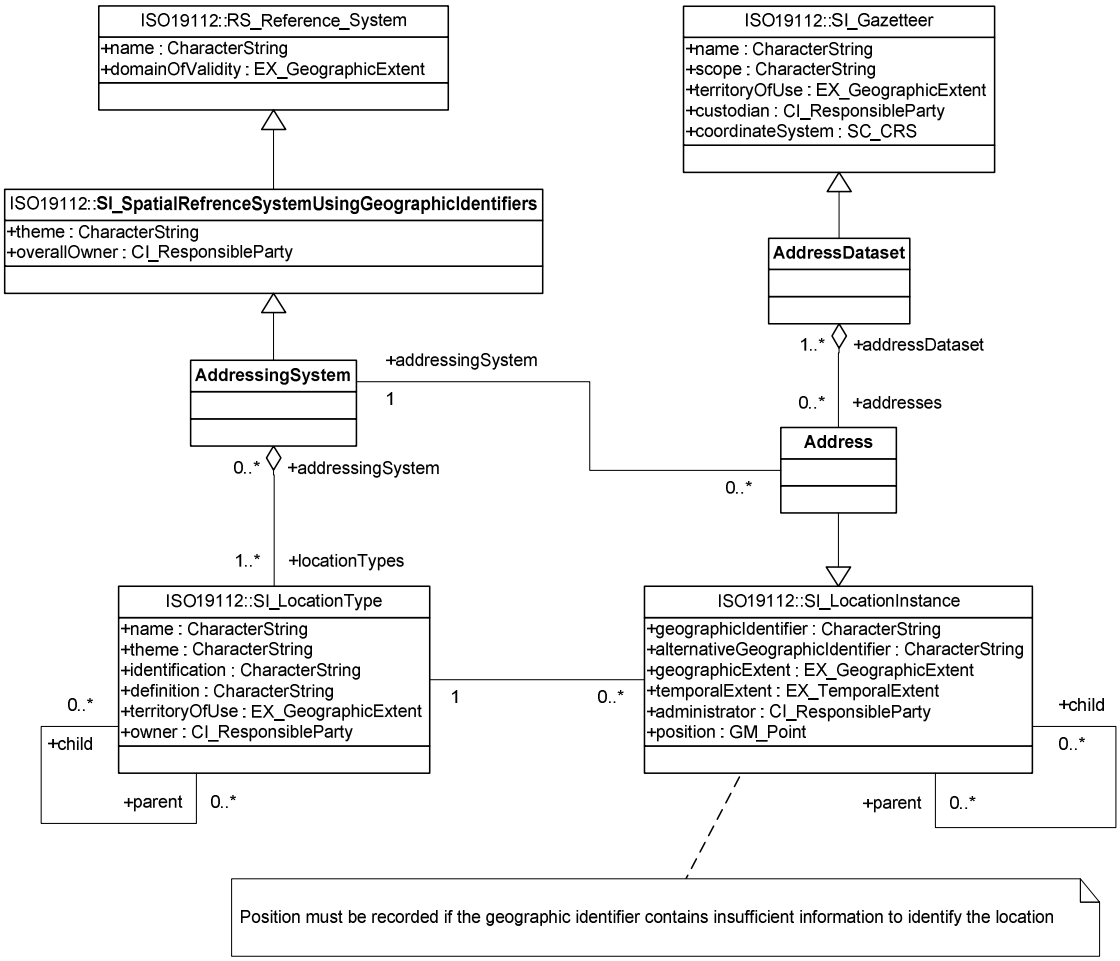


Figure 54. Data model for address data in Compartimos (adapted from ISO 19112:2003)

B.2 Data model: Address data catalogue

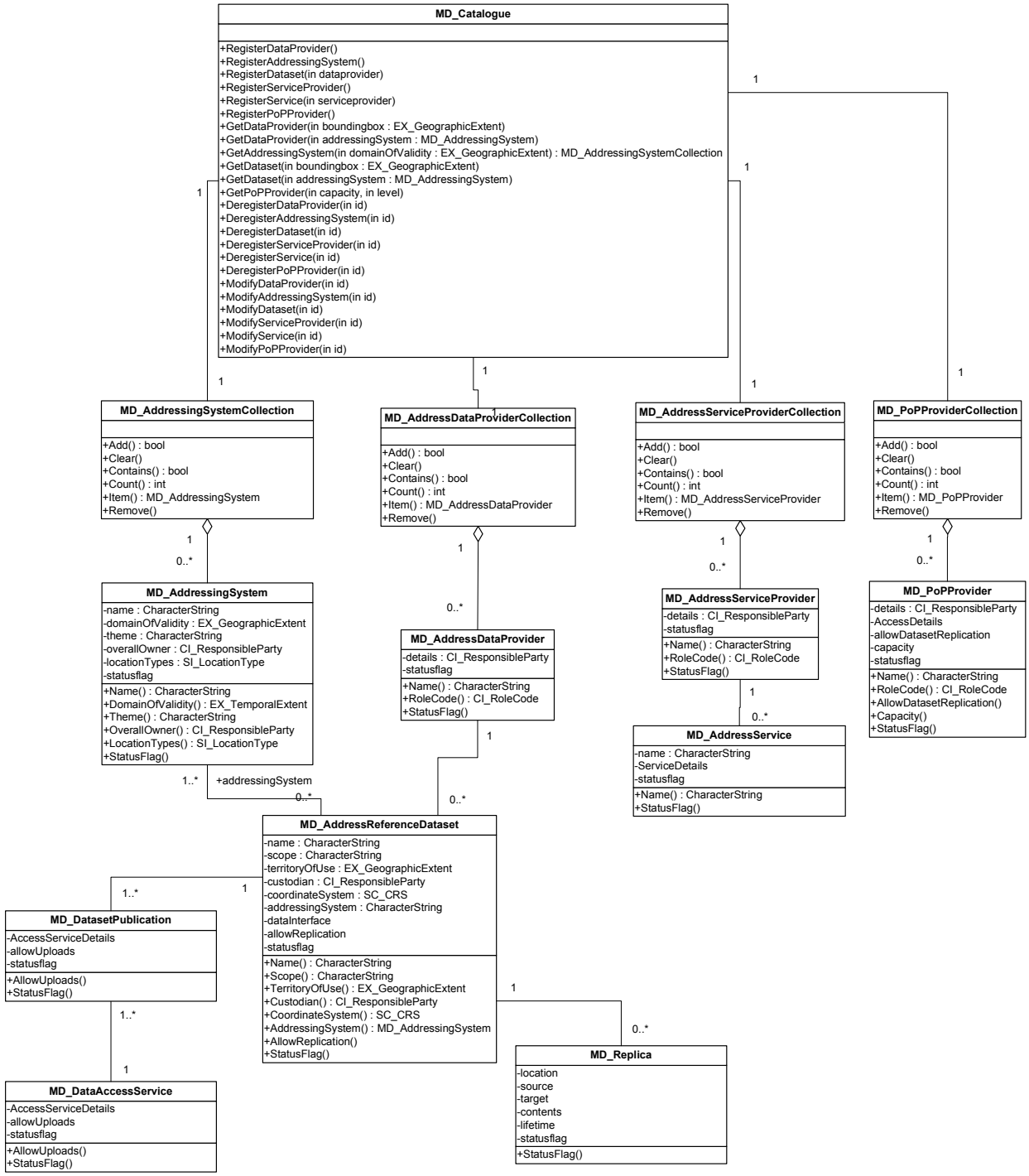


Figure 55. Data model: Address data catalogue

B.3 Sample data: Address data catalogue

lnk_AddressDataProvider_AddressDataset ...

dp	dsID
AfriGIS	AfriGIS NAD
AfriGIS	City of Tshwane street addresses
AfriGIS	AfriGIS Points of Interest

Record: 2 of 3

Figure 56. MD_AddressReferenceDataset

MD_AddressDataProvider : Table

dpID	rpOrgName	role	catStatus
1	AfriGIS	resourceProvider	Active
2	City of Tshwane Metropolitan Municipality	custodian	Active
3	The Knowledge Factory	resourceProvider	Inactive
(AutoNumber)			

Record: 1 of 3

Figure 57. MD_AddressDataProvider

MD_AddressDatasetPublication : Table

dPubID	dsID	name	allowUploads	status	adaID
1	AfriGIS NAD	Addresses for third party application developers	☐	Inactive	Intiendo web service
2	AfriGIS NAD	For municipal uploads	☑	Active	AfriGIS web service
3	AfriGIS Points of Interest	Addresses for third party application developers	☐	Inactive	Intiendo web service
(AutoNumber)			☐		

Record: 1 of 3

Figure 58. MD_DatasetPublication

MD_AddressingSystem : Table

RS_SpatialReferenceSystem_ID	name	domainOfValidity	theme	overAllOwner	versionNumber
1	Demarcation Board boundaries	South Africa	Province - District Municipality - Mur	Municipal Demarcation Board	
2	PAMSS post codes	South Africa	Post Office - Postcode	SA Post Office	
3	AfriGIS Street Address	South Africa	Province - Municipality - Town - Subu	AfriGIS	
4	AfriGIS Site Address	South Africa	Province - Municipality - Town - Subu	AfriGIS	
5	SANS 1883 street address type	South Africa		AfriGIS	
6	SANS 1883 site address type	South Africa		AfriGIS	
7	SANS 1883 landmark address type	South Africa		AfriGIS	
8	SANS 1883 intersection address type	South Africa		AfriGIS	
9	SANS 1883 building address type	South Africa		AfriGIS	
10	SANS 1883 SAPO box adress type	South Africa		SA Post Office	
11	SANS 1883 SAPO atreet address type	South Africa		SA Post Office	
12	SANS 1883 SAPO-type rural village address type	South Africa		SA Post Office	
13	SANS 1883 SAPO Poste restante address type	South Africa		SA Post Office	
(AutoNumber)					

Record: 1 of 13

Figure 59. MD_AddressingSystem

RS_SpatialReferenceSystemLocationTypes : Table

RS_SpatialReferenceSystem_ID	locationType	parent	child
Demarcation Board boundaries	Municipality	Province	Ward
Demarcation Board boundaries	District Municipality	Province	Municipality
Demarcation Board boundaries	Municipality	District Municipality	Ward
Demarcation Board boundaries	Ward	Municipality	None
Demarcation Board boundaries	Province	None	District Municipality
SANS 1883 SAPO box adress type	Province	None	Post Office
SANS 1883 SAPO box adress type	Post Office	Province	Postcode
SANS 1883 SAPO box adress type	Postcode	Post Office	None
SANS 1883 site address type	Municipality	Province	Town
SANS 1883 site address type	Town	Municipality	UsedName
SANS 1883 site address type	UsedName	Town	Site Number
SANS 1883 site address type	RegisteredName	Town	Site Number
SANS 1883 site address type	Site Number	RegisteredName	None
SANS 1883 site address type	RegisteredName	Municipality	Site Number
SANS 1883 site address type	UsedName	Municipality	Site Number
SANS 1883 site address type	Site Number	UsedName	None
SANS 1883 site address type	Province	None	Municipality
SANS 1883 street address type	CompleteAddressNı	CompleteStreetNam	None
SANS 1883 street address type	CompleteStreetNarr	RegisteredName	CompleteAddressN
SANS 1883 street address type	Province	None	Municipality
SANS 1883 street address type	Municipality	Province	Town
SANS 1883 street address type	Town	Municipality	UsedName
SANS 1883 street address type	CompleteStreetNarr	UsedName	CompleteAddressN
SANS 1883 street address type	UsedName	Municipality	CompleteStreetNar
SANS 1883 street address type	RegisteredName	Municipality	CompleteStreetNar
SANS 1883 street address type	RegisteredName	Town	CompleteStreetNar
SANS 1883 street address type	CompleteStreetNarr	None	None
SANS 1883 street address type	UsedName	Town	CompleteStreetNar
		None	None

Record: 1 of 28

Figure 60. MD_AddressingSystem.LocationTypes

SI_LocationType : Table

SI_LocationType_	name	theme	identification	definition	territoryOfUse	owner
-1	None	n/a	n/a	n/a		
1	Province	administrative	name	area	South Africa	Municipal Demarcation B
2	District Municipality	administrative	name | code	area	South Africa	Municipal Demarcation B
3	Municipality	administrative	name | code	area	South Africa	Municipal Demarcation B
4	Ward	administrative	number	area	South Africa	Municipal Demarcation B
5	Town	commonly known	name	area	South Africa	AfriGIS
6	UsedName	commonly known	name	area	South Africa	AfriGIS
7	RegisteredName	legal	name | code	area	South Africa	Chief Surveyor General
8	CompleteStreetName	as built	name	line	South Africa	AfriGIS
9	CompleteAddressNumber		number (text)	point	South Africa	AfriGIS
10	Site Number		number (text)	point	South Africa	AfriGIS
11	Post Office	postal	name	name only	South Africa	SA Post Office
12	Postcode	postal	code	code only	South Africa	SA Post Office
13	StreetNumberRange		numbers (text)	point	South Africa	AfriGIS
0						

Record: 2 of 14

Figure 61. SI_LocationType

Appendix C. Operations of the Compartimos service objects

C.1 CatalogueService

Table 40. Operations provided by the CatalogueService

Service name	Description
Start	Starts the catalogue service
Restart	Restarts the catalogue service
Stop	Stops the catalogue service
Synchronize	Synchronizes the local catalogue with the master catalogue
IsRunning	Returns true if the catalogue service has been started
DataPublications	Returns the collection of publications from the catalogue
Publish	Takes a dataset and data access service as input, and creates a dataset publication in the catalogue.
Update	Updates a dataset publication's information in the catalogue
Find	Finds a dataset publication(s) based on the specified input filter
Delete	Removes the dataset publication (i.e. the link between dataset and data access service) from the catalogue.
UsageInformation	Returns information about the usage of various datasets.
UpdateUsageInformation	Updates dataset usage information in the catalogue. This information is used by the ReplicaService to decide when to replicate a dataset.
About an addressing system	
AddressingSystems	Returns the collection of addressing systems from in the catalogue
RegisterAddressingSystem	Adds information about an addressing system to the catalogue.
UpdateAddressingSystem	Updates addressing system information in the catalogue. Fails if the addressing system is associated with a dataset in the catalogue.
FindAddressingSystem	Finds an addressing system(s) based on the specified input filer
DeleteAddressingSystem	Removes the addressing system information from the catalogue. Fails if the addressing system is still associated with a dataset in the catalogue.
About an address-related service	
Services	Returns the services provided by a specific service provider
RegisterService	Adds information about an address-related service to the catalogue.
UpdateService	Updates the information about an address-related service's in the catalogue.
FindService	Finds a service(s) based on the specified input filter
DeleteService	Removes the information about the address-related service from the catalogue.

Service name	Description
About a data provider	
DataProviders	Returns the collection of data providers from the catalogue
RegisterDataProvider	Adds a data provider's information to the catalogue.
UpdateDataProvider	Updates a data provider's information in the catalogue.
FindDataProvider	Finds a data provider(s) based on the specified input filer
DeleteDataProvider	Deletes a data provider from the catalogue
About a data access service	
AddressDataAccessServices	Returns the collection of address data access service from the catalogue
RegisterDataAccessService	Adds information about an address data access service to the catalogue.
UpdateDataAccessService	Updates information about the data access service in the catalogue.
FindDataAccessService	Finds a data access service(s) based on the specified input filter
DeleteDataAccessService	Removes the information about the access service from the catalogue Fails if a dataset is associated with the service through a publication.
About a dataset	
Datasets	Returns the collection of datasets from the catalogue
RegisterDataset	Adds information about an address dataset to the catalogue.
UpdateDataset	Updates a dataset's information in the catalogue.
FindDataset	Finds a dataset(s) based on the specified input filter
DeleteDataset	Removes the dataset's information from the catalogue Fails if a data access service is associated with the dataset through a publication.
About a node host	
NodeHosts	Returns the collection of node hosts from the catalogue
RegisterNodeHost	Adds information about a node host to the catalogue.
UpdateodeHost	Updates a node host's information in the catalogue.
FindNodeHost	Finds a node host(s) based on the specified input filter
DeleteNodeHost	Removes information about the node host from the catalogue.
About a replica	
Replicas	Returns the collection of replicas from the catalogue
RegisterReplica	Adds replica information to the catalogue
UpdateReplica	Updates a replica's information in the catalogue
FindReplica	Finds a replica(s) based on the specified input filter
DeleteReplica	Delete a replica's information from the catalogue
About a service provider	
ServiceProviders	Returns the collection of service providers from the catalogue
RegisterServiceProvider	Adds a service provider's information to the catalogue.
UpdateServiceProvider	Updates a service provider's information in the catalogue.
FindServicProvider	Finds a service provider(s) based on the specified input filter
DeleteServiceProvider	Removes the information about the service provider from the catalogue. Fails if the service provider is still associated to an address-related service in the catalogue.

C.2 ReplicaService

Table 41. Services provided by the ReplicaService

Service name	Description
CreateReplica	Creates a replica either of the complete dataset, or of the higher-level location type values only, depending on the parameters. The relevant replica information is updated in the catalogue.
ValidateReplica	Determines whether the replica is identical to the primary source from where it has been replicated
ModifyReplicaContents	Change the set of data that is being replicated, e.g. by specifying specific higher-level location types of an addressing system
DeleteReplica	Removes the replica from the node, and also deletes the associated information from the catalogue.
SynchronizeReplica	Makes use of the transfer service to update a replica with the changes that occurred at its dataset.

C.3 TransferService

Table 42. Operations provided by the TransferService

Service name	Description
SetupTransfer	Initializes a transfer between two locations.
GetTransferState	Returns information about the state of a transfer
StartTransfer	Starts transferring the data as specified in the SetupTransfer service
PauseTransfer	Pauses the transfer, i.e. data that has been transferred does not have to be transferred again when the transfer is resumed.
ResumeTransfer	Resumes the paused transfer.
StopTransfer	Stops the transfer, i.e. any data that has been transferred until that time, is lost. The transfer has to be started again with the StartTransfer service.

C.4 AddressDataAccessService

Table 43. Services by the AddressDataAccessService

Service name	Description
CreateAddressDataset	Takes addressing systems as input parameters and physically creates a database that can store addresses of those addressing systems.
GetAddress	Returns the specified addresses
UploadAddressData	Uploads a whole dataset, or large amount of data, into the address dataset (used for replication)
AddAddress	Adds a single address to the address dataset (used for replication)

C.5 VirtualAddressDataService

Table 44. Services by the VirtualAddressDataService

Service name	Description
GetAddress	Returns the specified addresses. Depending on the input parameters the results are either returned as output parameters, or the TransferService is employed to transfer a file containing the resulting data.
UploadAddressData	Uploads address data in bulk into the address data grid. The uploaded data is stored as a single dataset at any data provider with a data access service for which the AllowUploads property is true.

Appendix D. Additional Compartimos use cases

D.1 Upload an address dataset

This use case illustrates the interaction of objects in the address data grid when a data provider uploads data into the address data grid that will be hosted by a third party data host. This happens, for example, when a data provider such as a small local authority that does not have the resources to host a dataset, makes use of a third party to host its dataset. Such a data provider will update its dataset once, and subsequently upload either re-upload the dataset or upload additions and modifications to the dataset. The TransferService is employed to upload the dataset as a single file.

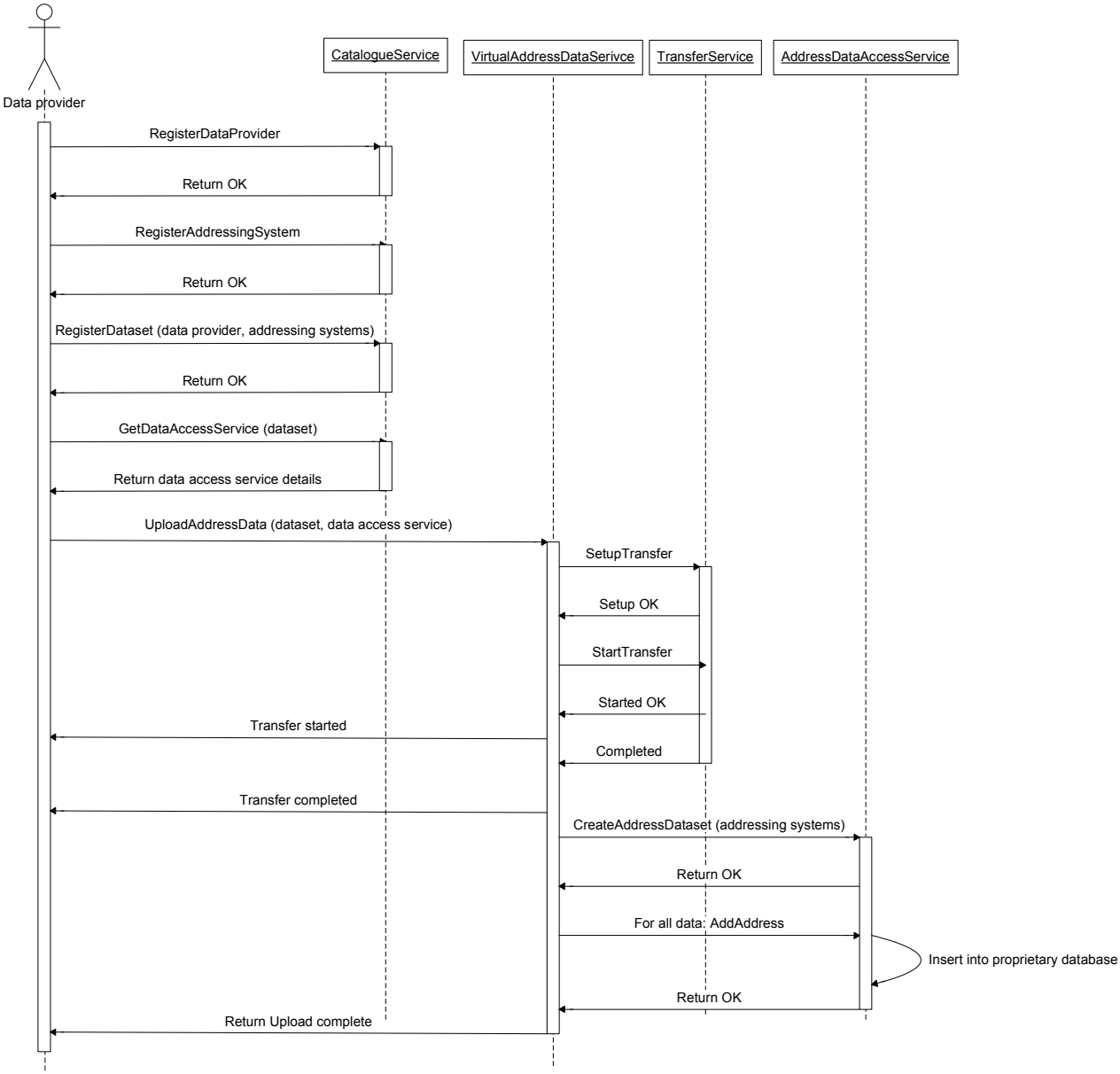


Figure 62. Sequence diagram for uploading an address dataset

D.2 Address dataset publication on the grid

In order to make a dataset available in the address data grid, the following has to happen:

1. Register a data provider in the metadata catalogue through the CatalogueService.
2. Ensure that the addressing systems required for the dataset are in the catalogue; otherwise register the required addressing systems through the CatalogueService.
3. Register the data access service that will provide uniform access to the dataset using the CatalogueService.
4. Register the dataset in the catalogue using the CatalogueService.
5. Register a data publication in the catalogue by associating the dataset and data access service in the catalogue.

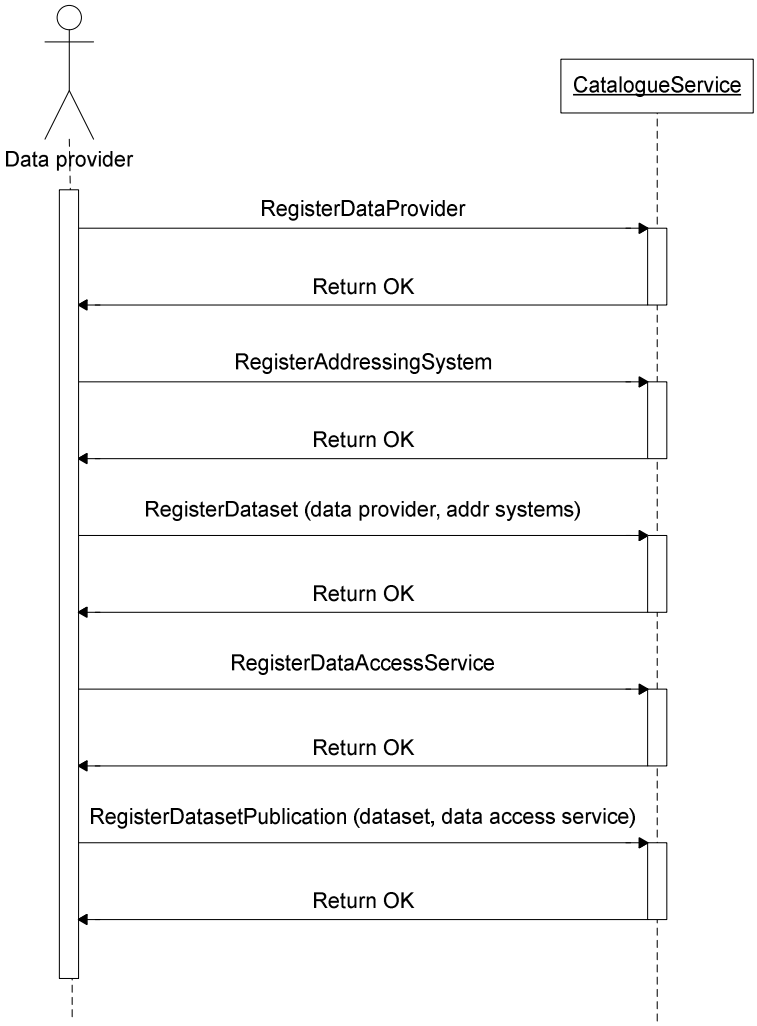


Figure 63. Sequence diagram for address dataset publication on the grid

D.3 Publication of an address-related service on the grid

To publish an address-related service on the address data grid requires two steps only:

1. Register a service provider in the catalogue using the CatalogueService.
2. Register the address-related service in the catalogue using the CatalogueService.

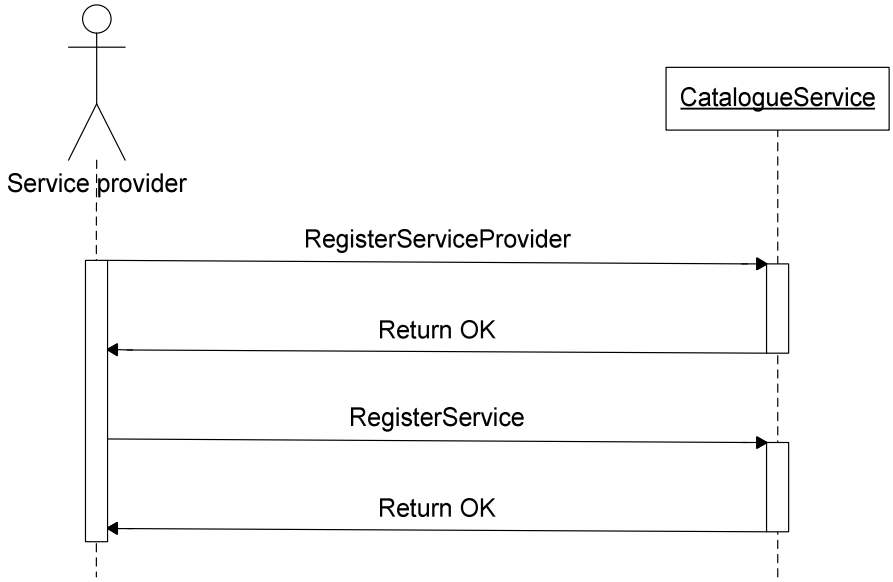


Figure 64. Sequence diagram for publication of an address-related service on the grid

D.4 Dataset replication

The following sequence diagram provides a simplified view of dataset replication. The dataset usage information is kept up to date by the VirtualAddressDataService whenever it gets a request for data. The ReplicaService polls this information at regular intervals and based on the thresholds set in the catalogue decides when it is necessary to replicate a dataset. Note that a dataset can only be replicated if it has been registered in the catalogue as a dataset that allows replication. As part of the ReplicaService's *CreateReplica*, the required information is set up in the catalogue, and upon completion the replica is synchronized, i.e. it is copied from source to target making use of the TransferService.

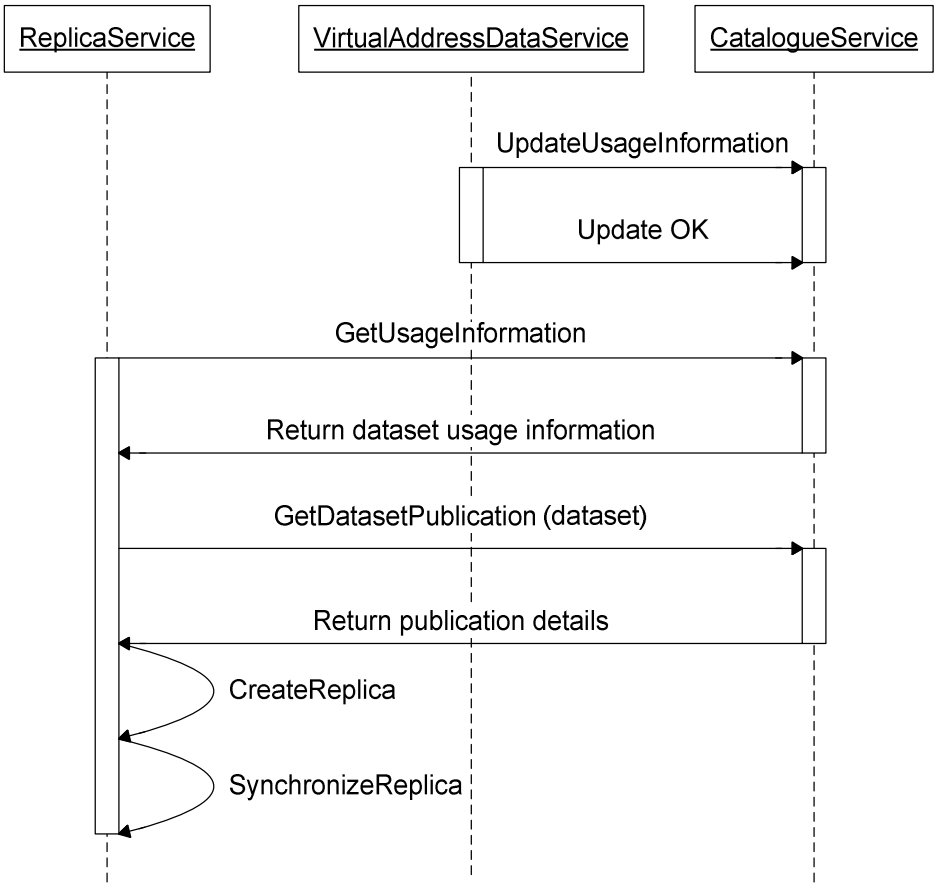


Figure 65. Sequence diagram for dataset replication

Appendix E. Journal publications

E.1 Address databases for national SDI: Comparing the novel data grid approach to data harvesting and federated databases

Coetzee S and Bishop J (2008). Address databases for national SDI: Comparing the novel data grid approach to data harvesting and federated databases, *International Journal of Geographic Information Science*, 26 September 2008, available online ahead of print edition at http://www.tandf.co.uk/journals/tf/13658816.html, accessed 13 November 2008.

The complete paper is included as Chapter 6 of the dissertation. This appendix includes only the cover pages.

E.2 What is an address in South Africa?

Coetzee S and Cooper AK (2007b). What is an address in South Africa? *South African Journal of Science*, Nov/Dec 2007, **103**(11/12), pp449-458.