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?

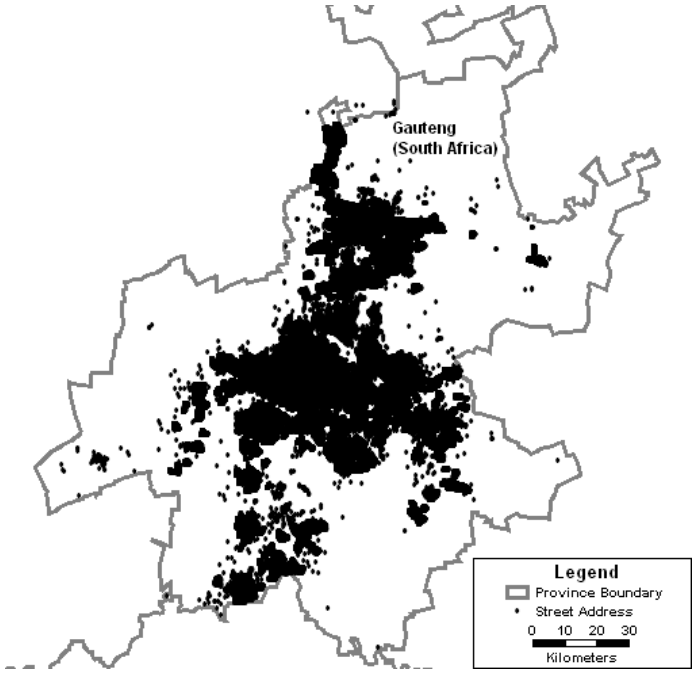

Figure 43. Street addresses in Gauteng (Source: AfriGIS NAD)

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.

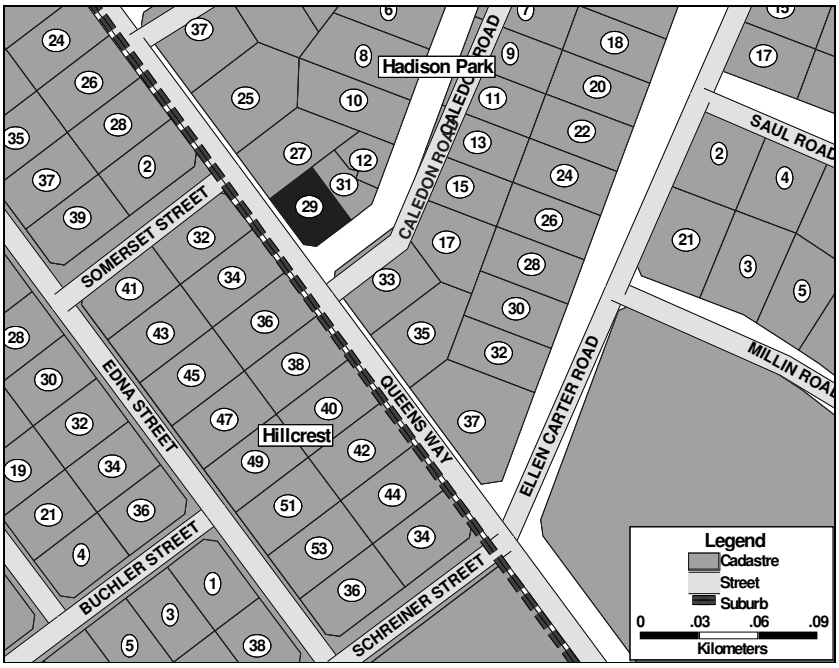

Figure 45. Hillcrest and Hadison Park in Kimberley (Source: AfriGIS NAD)

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.

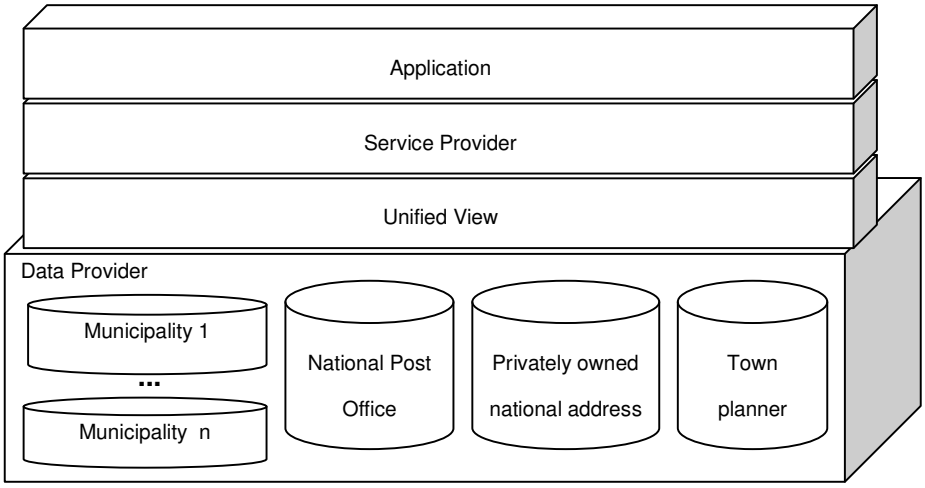

Figure 46. National address database

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.

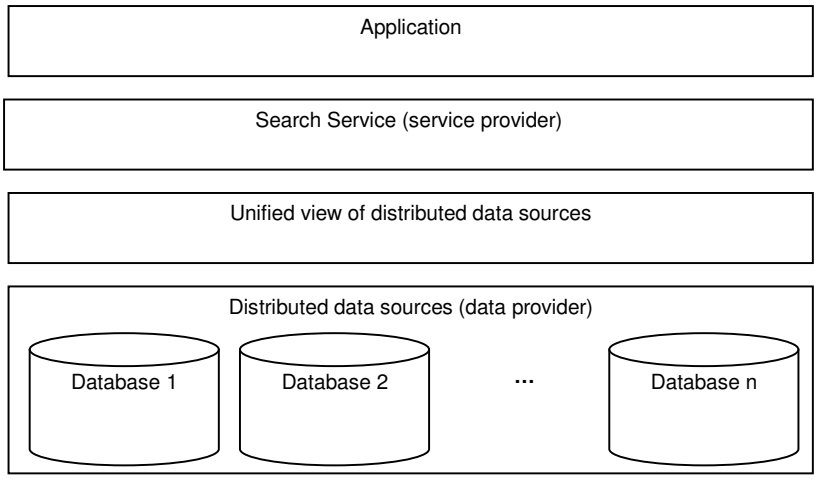

Figure 47. Information federation models

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.

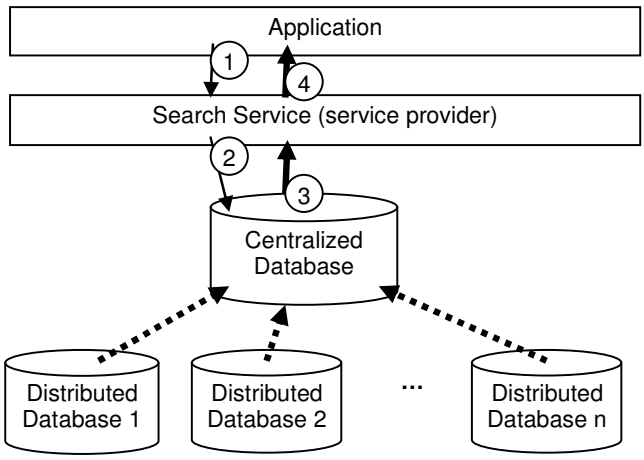

Figure 48. The data harvesting model

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.

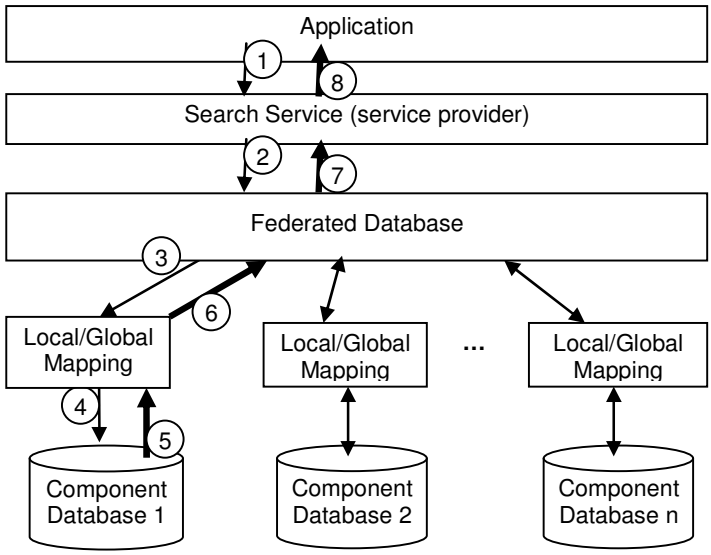

Figure 49. The federated database model

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).

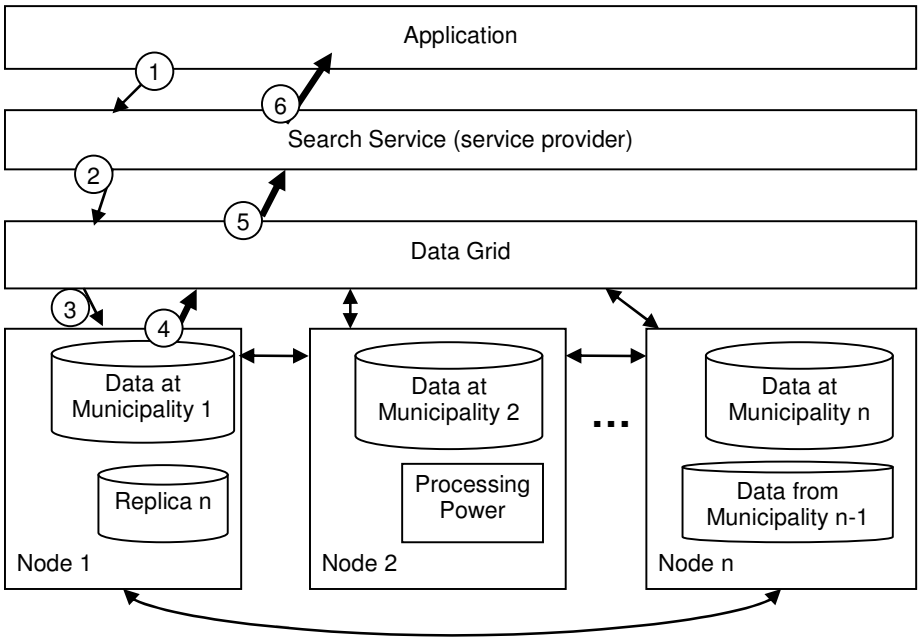

Figure 50. The data grid model

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.

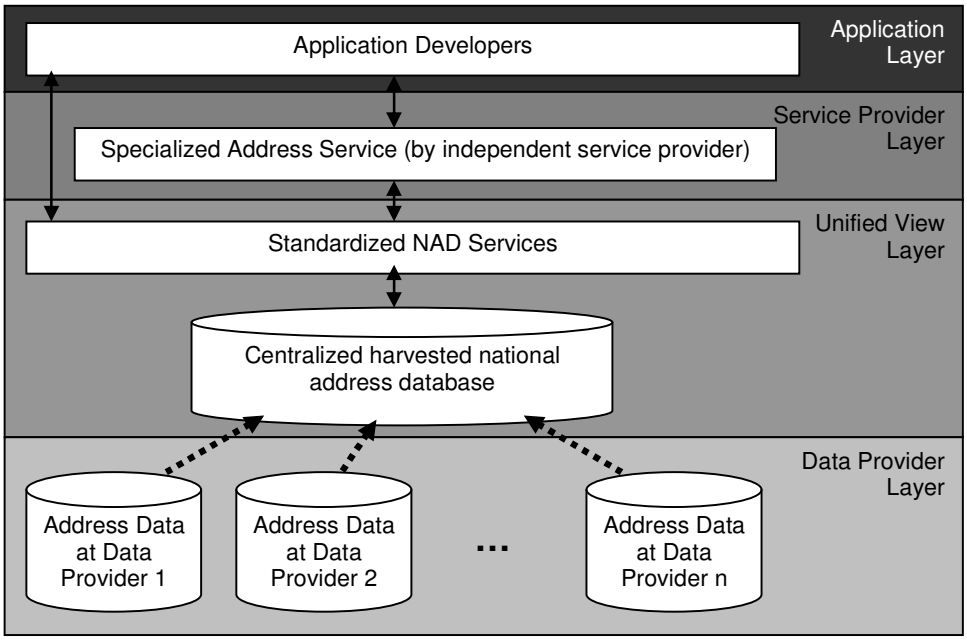

Figure 51. Single centralized harvested national address database

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.

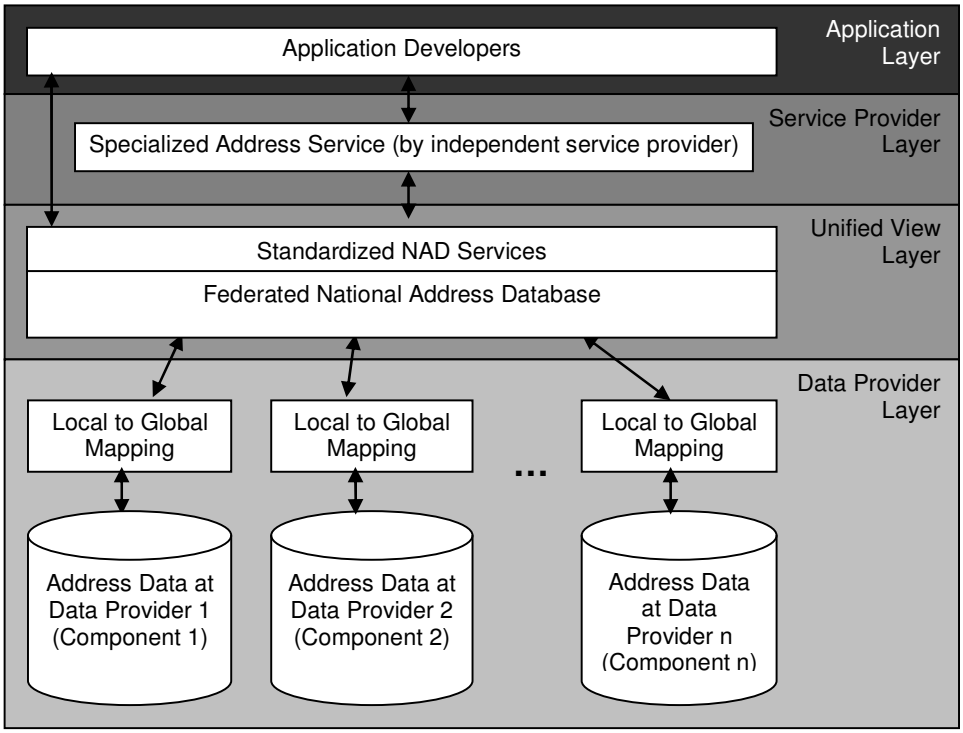

Figure 52. Federated national address database

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.

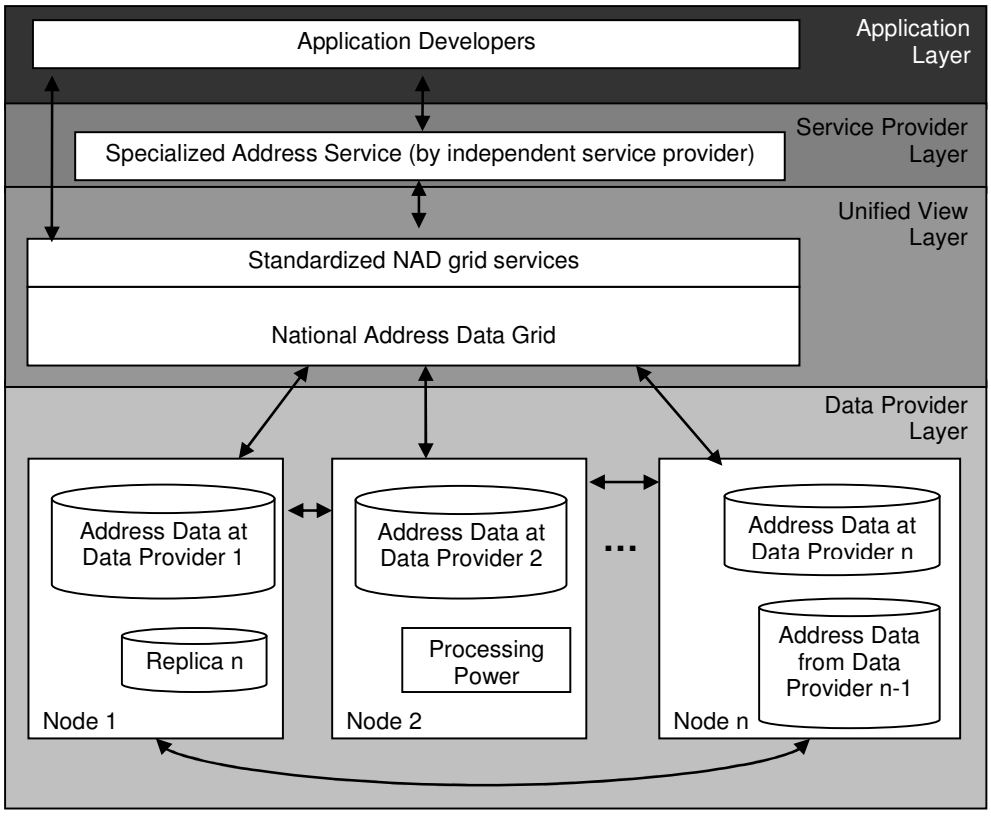

Figure 53. The national address database as a data grid

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.