

Chapter 5 Implementation and evaluation of Compartimos

5.1 Introduction

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

5.2 Technology choices of specific Compartimos objects

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

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

Table 13. Overview of technology choices for Compartimos objects

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

5.2.1 The catalogue

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

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

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

The Compartimos catalogue includes two types of metadata:

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

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

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

5.2.2 The catalogue service (CatalogueService)

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

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

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

5.2.3 The replica service (ReplicaService)

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

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

5.2.4 The transfer service (TransferService)

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

5.2.5 The address data access service (AddressDataAccessService)

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

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

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

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

5.2.6 The virtual address data service (VirtualAddressDataService)

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

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

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

5.2.7 The address dataset (AddressDataset)

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

5.2.8 The address-related service (AddressService)

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

5.3 Overall technology choices for Compartimos

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

5.3.1 Security

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

5.3.2 Operating system and/or programming language

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

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

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

Table 14. Where Compartimos services are hosted

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

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

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

5.3.3 Web service protocols

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

The choice of communication protocol is also related to available standards. The OGF, where grid standardization takes place, are cooperating with OASIS and thus web services based on SOAP are prevalent. The OGC Web Service Implementation Specifications on the other hand, is based on POST/GET methods through HTTP. This poses problems for any integration of OGF and OGC web services. However, a recently completed OGC Web Services, Phase 5 (OWS-5) Testbed, included the development of SOAP and WSDL interfaces for four services: WMS, WFS-T, WCS-T, and WPS (OGC 2008b), showing that OGC is paying attention to this matter.

5.3.4 Connectivity and bandwidth

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

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

5.4 Proof of concept implementation of Compartimos

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

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

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

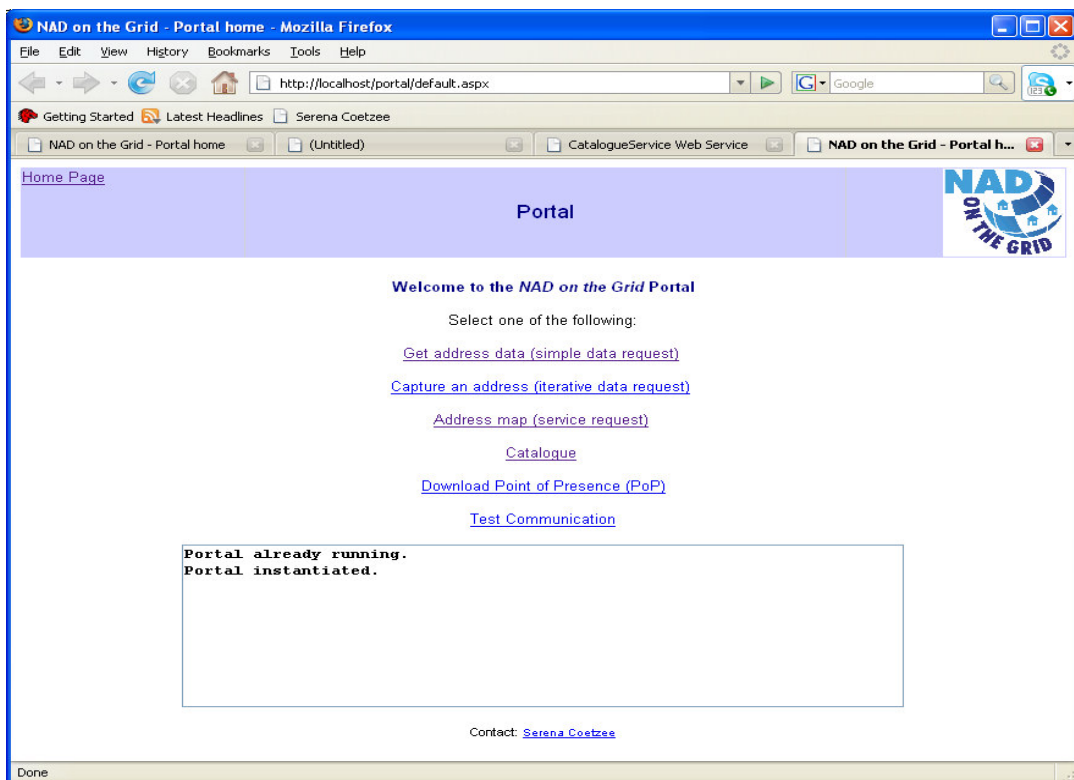
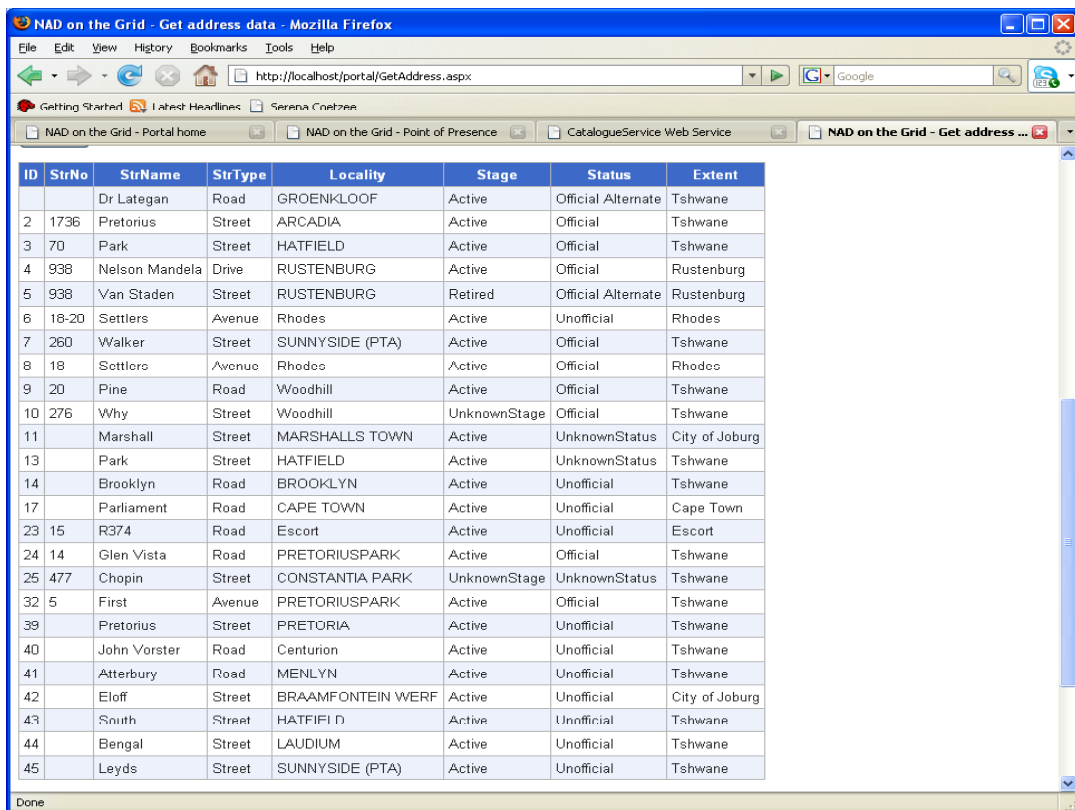


Figure 40. Home page of the address data grid portal

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

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



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

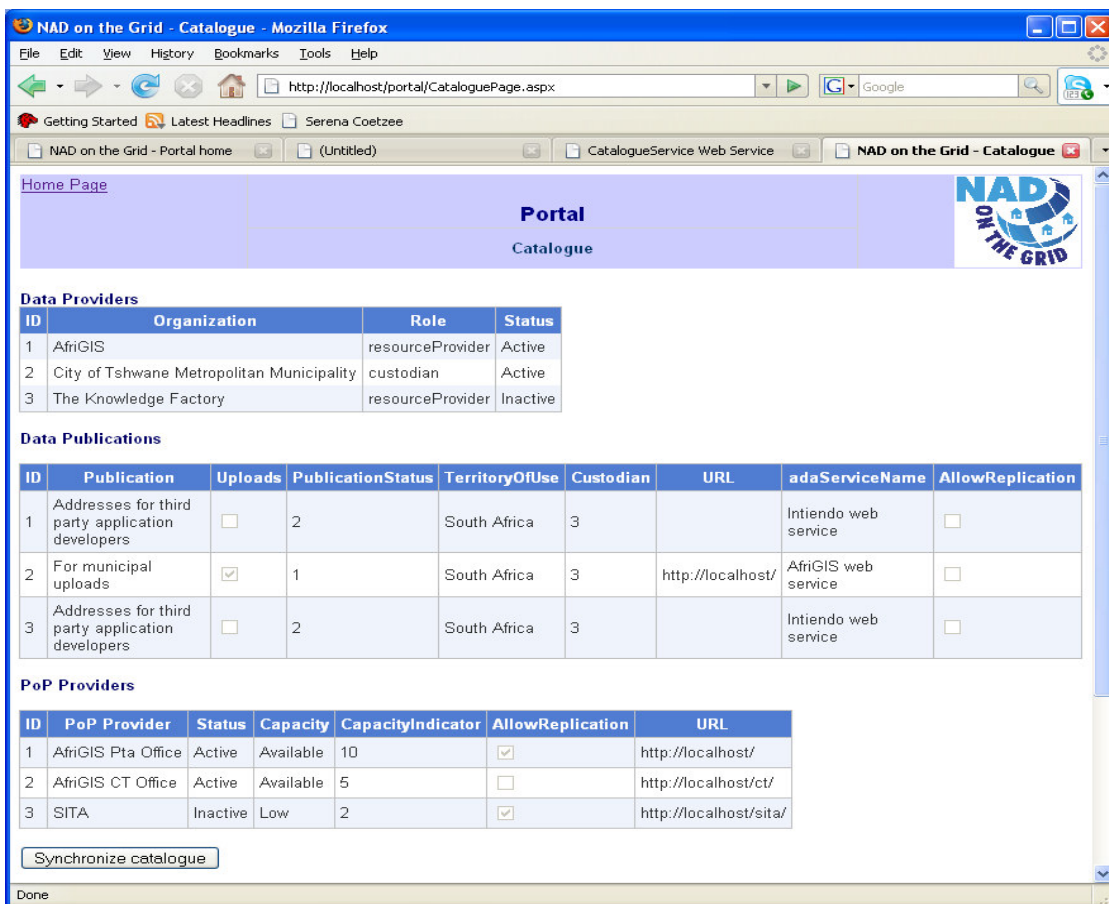
Figure 41. Results of a simple data request displayed in the portal

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

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

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

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



Data Providers

ID	Organization	Role	Status
1	AfriGIS	resourceProvider	Active
2	City of Tshwane Metropolitan Municipality	custodian	Active
3	The Knowledge Factory	resourceProvider	Inactive

Data Publications

ID	Publication	Uploads	PublicationStatus	TerritoryOfUse	Custodian	URL	adaServiceName	AllowReplication
1	Addresses for third party application developers	<input type="checkbox"/>	2	South Africa	3		Intiendo web service	<input type="checkbox"/>
2	For municipal uploads	<input checked="" type="checkbox"/>	1	South Africa	3	http://localhost/	AfriGIS web service	<input type="checkbox"/>
3	Addresses for third party application developers	<input type="checkbox"/>	2	South Africa	3		Intiendo web service	<input type="checkbox"/>

PoP Providers

ID	PoP Provider	Status	Capacity	CapacityIndicator	AllowReplication	URL
1	AfriGIS Pta Office	Active	Available	10	<input checked="" type="checkbox"/>	http://localhost/
2	AfriGIS CT Office	Active	Available	5	<input type="checkbox"/>	http://localhost/ct/
3	SITA	Inactive	Low	2	<input checked="" type="checkbox"/>	http://localhost/sita/

Synchronize catalogue

Figure 42. Catalogue contents displayed in the portal

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

5.5 Evaluation of Compartimos

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

5.5.1 Evaluation against criteria of the evaluation framework

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

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

Table 15. Infrastructure

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

Table 16. Data providers

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

Table 17. Naming

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

Table 18. Address Dynamics

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

Table 19. Accessibility

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

Table 20. Security

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

Table 21. Organizational Issues

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

5.5.2 Discussion of results

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

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

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

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

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

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

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

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

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