Reference model for a data grid approach to address data in a dynamic SDI

Serena Coetzee

Abstract A grid is concerned with the integration, virtualization, and management of services and resources in a distributed, heterogeneous environment that supports virtual organizations across traditional administrative and organizational domains. Spatial data infrastructures (SDI) aim to make spatial data from multiple sources available and usable to as wide an audience as possible. The first SDIs of the 1990s followed a top-down approach with the focus on data production and centralization. In recent years, SDIs have seen a huge increase in the number of participants, necessitating a more dynamic bottom-up approach. While much research has been done on web services and SDIs, research on the use of data grids for SDIs is limited. In this paper an emergency response scenario is presented to illustrate how the data grid approach can be used as decentralized platform for address data in a dynamic SDI. Next, Compartimos (Spanish for 'we share') is presented, a reference model for an address data grid in an SDI based on the Open Grid Services Architecture (OGSA). Compartimos identifies the essential components and their capabilities required for a decentralized address data grid in a dynamic SDI. It deviates from the current centralized approach, allows data resources to come and go and node hosts to grow and shrink as necessary. An address data grid in an SDI is both a novel application for data grids as well as a novel technology in SDI environments and thus advances the mutual understanding between data grids and SDIs. In conclusion, additional research required for address data grids in SDIs is discussed.

Keywords geospatial data · data grid · spatial data infrastructure · SDI · address data · data sharing

1 Introduction

A grid is a system that is concerned with the integration, virtualization, and management of services and resources in a distributed, heterogeneous environment that supports virtual organizations (collections of users and resources) across traditional administrative and organizational domains (real organizations) [1]. How virtual organizations collaborate and share resources in order to achieve a common goal is described as the 'grid architecture' in *The Anatomy of the grid* [2] and *The Physiology of the grid* [3]. This has subsequently evolved into the Open Grid Services Architecture (OGSA) published by the Open Grid Forum (OGF) [4], a vision of a broadly applicable and adopted framework for grids.

A data grid is a special kind of grid in which data resources are shared and coordinated. The OGSA data architecture [5] describes the interfaces, behaviors and bindings for manipulating data within the broader OGSA. It presents a "toolkit" of data services and interfaces that can be composed in a variety of ways to address multiple scenarios. These services and interfaces include data access, data transfer, storage management, data replication, data caching, and data

S. Coetzee (✉)
IT Building 4-38, Department of Computer Science, University of Pretoria, Pretoria, 0002, South Africa.
Tel: +27 12 420 2547 · Fax: +27 12 362 5188 · E-mail: serena.coetzee@up.ac.za

federation. The components of the data architecture can be put together to build a wide variety of solutions.

A spatial data infrastructure (SDI) aims to make spatial data usable by people. Technologies, systems (hardware and software), standards, policies, agreements, human and economic resources, institutions and organizational aspects have to be carefully orchestrated to make this possible. National SDIs emerged in the early 1980s in countries such as the USA and Australia. These first generation SDIs mostly followed a top-down product-based approach that was centrally coordinated, typically by a national mapping agency. The next generation of SDIs followed a more process-based approach focusing on the creation of a suitable infrastructure to facilitate the management of information access (mostly read-only), instead of the linkage to existing and future databases [6]. Web services are a prominent feature of process-based SDIs and in this generation of SDIs the trend towards decentralized and distributed networks, similar to the Internet and World Wide Web, started. Today, SDIs are evolving to accommodate the challenges of ubiquitous read-write access by millions of users from all kinds of devices. The strict top-down approach of early SDIs is evolving into more dynamic bottom-up approaches. Masser *et al.* [7] 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.

Due to their service, infrastructure and land administration responsibilities, it is commonly found that it is the local authority that establishes and maintains address data for its area of jurisdiction [8,9]. However, address data is often required at a larger scale, for example, for the planning and management of national Census and election operations. The principles of SDIs therefore apply to address data. Currently, many national address databases of the world follow the centralized approach where address data is loaded into a single centralized database [10,11,12]. This approach implies that there is a single entity that maintains the centralized database, either because it has a public mandate or for commercial gain. However, the demand for address data is changing along with the evolution of the Web. In true Web 2.0 style, users want to not only view addresses on a map, but want to suggest a correction to an address from their mobile phone, contribute a new address from a handheld GPS device or use address data for routing in their in-vehicle navigation system. Instead of a single static centralized database, ubiquitous read-write access by millions of users from all kinds of devices is required.

There is an abundance of definitions for a grid, but one commonly cited definition is Foster's [13] three point check list, stating that a grid is a system that (1) coordinates resources that are *not subject to centralized control*; (2) delivers *non-trivial qualities of service*; and (3) uses *standard, open, general-purpose protocols and interfaces*. Looking at the evolving requirements for address data in an SDI, there are some similarities with a data grid to be found. A definite requirement for address data in an SDI is *non-trivial qualities of service*, such as simple data retrieval services but also update/edit services and more sophisticated address-related services such as geocoding, verification, routing and mapping. Included in the latter is the requirement for scalability to cater for ubiquitous read-write access from millions of users and for high volume data transfers. The heterogeneous address data *resources* are distributed among local authorities, as well as independent organizations with datasets to which users can contribute address data, known as volunteered geographic information (VGI). These resources have to be

coordinated into a national dataset and are *not subject to centralized control*. Because of the heterogeneous environment and the requirement for ubiquitous access, *standard, open protocols and interfaces* are imperative. The similarities between a data grid and address data in an SDI are thus clearly evident.

The work reported in this paper is part of a research project on 'Distributed Address Management' with the objective of investigating the data grid approach and its design imperatives for national address databases in an SDI. In earlier work, a novel evaluation framework for national address databases was used to evaluate the data grid approach and other information federation models for the use in address databases for national SDI. The data grid approach deviates from the centralized approach in the other models and this makes it suitable for SDIs that are currently evolving towards more distribution and more dynamism. The evaluation showed that where a large number of organizations with multiple heterogeneous address datasets are involved with no single organization tasked with the management of a national address database, the data grid approach has some unique features that make it an attractive alternative to the other models [14].

As a next step in our research, Compartimos, a reference model for a data grid architecture for address data, was developed. A reference model is an abstract framework for understanding significant relationships among the entities of some environment (OASIS 2008). The Compartimos reference model provides such an abstract representation of the essential components and their relationships that are required for an address data grid in an SDI environment. Compartimos serves to analyze the problem space of data grids and SDIs by addressing a very specific problem in these areas (address data on a data grid in an SDI) and provides valuable feedback about the usability of the general models (data grids for address data in an SDI) in this specific area of interest. Compartimos also has a clarifying purpose, because it is one of the first attempts in the joint SDI and data grid problem space, and thus enhances the mutual understanding of these two domains. The two problem spaces respectively represent two disciplines: the data grid problem space in the Computer Science discipline and the SDI problem space in the Geographic Information Science discipline.

In this paper we present Compartimos and what we learnt from its design about the challenges of grid-enabling address data in an SDI. The objectives of this paper are to present 2) an emergency response scenario that illustrates how the data grid approach can be used for address data in an SDI; 3) Compartimos, a reference model for an address data grid in an SDI based on the OGSA data architecture; 4) a discussion of the lessons learnt from Compartimos and challenges for grid-enabling address data in an SDI; and 5) a conclusion and future research directions in this area.

The introduction is concluded here with a short discussion of related work to illustrate the novelty of our work. Reports on grid computing for spatial data in general are found in Hua *et al.* [15], Aloisio *et al.* [16], Allen *et al.* [17], Aydin *et al.* [18] and Xue *et al.* [19]. First research reports on grid computing technologies in SDI environments are found in the papers by Aloisio *et al.* [20], Shu *et al.* [21], Wei *et al.* [22], Padberg and Kiehle [23], Yue *et al.* [24], Hui *et al.* [25], as well as the Geodateninfrastruktur-grid (GDI-grid) project (http://www.dgrid.de/index.php ?id=398&L=1), which is part of D-grid, a long-term German strategic initiative in grid computing. The Gis.Science journal [26] recently featured a special issue on 'Grid computing

and GIS'. A few of the more recent reports investigate grid computing (as opposed to data grids) in an SDI; none of them reports on address data.

The initial focus of the Memorandum of Understanding (MoU) between the Open Geospatial Consortium (OGC) and the OGF [27] 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 UK's National Grid Service. Results from the work reported on in this paper suggest that grid-enabling spatial data integration in an SDI environment should also be explored, i.e. grid-enabling other web services specified by OGC, such as the Web Feature Service (WFS). The OGC-OGF collaboration proves that the international geospatial community is increasingly interested in utilizing grid technology as a solution to its problems, while the grid community has found another user community that can benefit from its technology.

In the position paper by Craglia *et al.* [28], a group of international geographic and environmental scientists from government, industry and academia present the vision of the next generation Digital Earth and identify priority research areas to support this vision. These include information integration and computational infrastructures, also addressed in the research reported here.

The related work confirms that the data grid approach for address data in an SDI is innovative and new, and it proves that the work is relevant at this point in time, both in Computer Science (where grid computing is studied) and in Geographic Information Science (where SDIs are studied). The data grid approach to address data in an SDI is both a novel application of data grids as well as a novel technology in SDIs. This work is unique because Compartimos is designed for *address data* in a data grid.

2 Emergency response scenario

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

To map the location of damage or distress reports, the address on an incoming report is matched to an address in a reference address dataset that includes geo-spatial coordinates, a process known as geocoding. The ERC is in possession of software that automates the geocoding but this software requires the address data to be in a single database, structured according to a specific data model. The address data is further used as backdrop for any maps that are sent to the

rescue and clean-up teams. Address data has to be collected from the 50 odd individual cities and towns that have been affected by the storm. Each of these datasets might have a different model, format and licensing agreement. For some areas, no address data exists at all.

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

The first option is to force the heterogeneous address data from the various cities into a single uniform database as required for the geocoding tool. Any address data that cannot be converted into the uniform data model is lost. A second option is to set-up the geocoding tool to work for each of the different data formats from the individual cities and towns, i.e. fifty different configurations of the geocoding software in the worst case. A third option is to add individual datasets to one large map on which geocoding is done manually by humans, a time consuming process!

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

Applications, processing cycles and datasets are abstracted as resources in a grid world. Each resource can be accessed remotely according to its individual policy. Thus each city can securely grant rights to the ERC for access to its address dataset. This eliminates the need to download data or physically collect data, while at the same time protecting the privacy and integrity of the city's data. In addition, the ERC, or any other organization participating in the relief, can host an address data resource to which any user 'out there' can contribute address data, i.e. address data as VGI. In this way, address data for areas where it does not yet exist, can be captured.

An address data grid also requires standardization in terms of address data exchange. Even though each city maintains its data according to its own data model, it publishes and makes available a grid-enabled web service, or a grid service, that provides access to its address data according to an agreed upon address data exchange standard and protocol. The geocoding software makes use of these grid services to seamlessly work with the data from any city. In this way the ERC is guaranteed to display the latest address data on the map. Fig. 1 shows how the different components interact in this scenario.

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

Further, if the geocoding software is grid-enabled, it can execute in parallel on the grid processing resources of the different cities. Naturally, when disaster strikes, it does not affect a single address location but an area comprising numerous address locations. Thus an alternative strategy would be the following: when a geocoding request is received for a specific city's address data, the address reference for that suburb and its neighborhood is immediately replicated at the ERC. Subsequent geocoding requests from that area are then processed locally (and therefore faster) at the ERC.

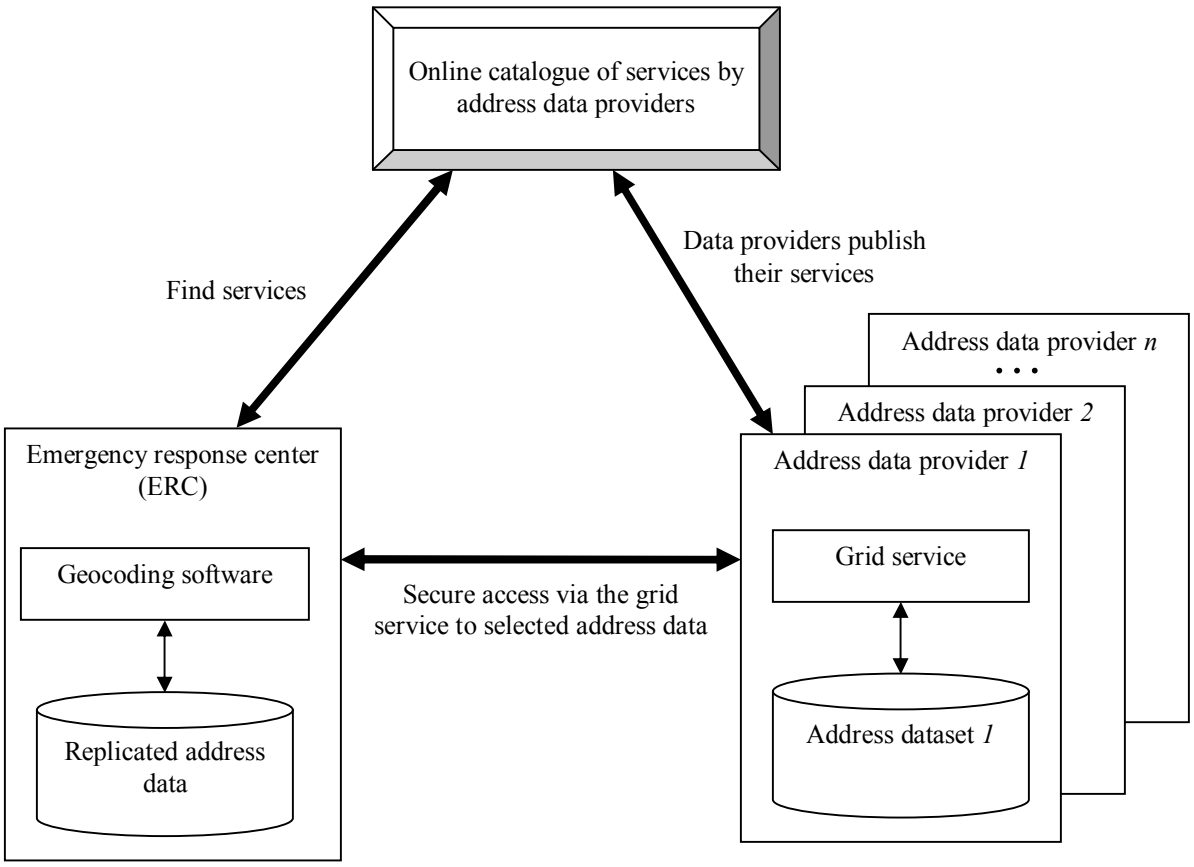

Fig. 1 Mapping the locations of damage and distress reports

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

3 The Compartimos reference model

In this section Compartimos, a reference model for an address data grid in an SDI based on the OGSA data architecture, is presented. Section 3.1 relates address data in an SDI to the OGSA data architecture. A virtual organization (VO) comprises the set of individuals and/or institutions sharing resources in a grid and in section 3.2 the VO for the specific case of an address data grid in an SDI is described. The Compartimos components, which are the essential data grid components that interact at interfaces, enabling the single virtual address dataset, are described in section 3.3. Each component is compared to its 'pure' OGSA counterpart, highlighting the address- and SDI-specific capabilities. The OGSA data architecture describes security issues that

are important in a data grid, which are equally applicable to address data in an SDI and these are discussed in section 3.4.

3.1 Address data in an SDI and the OGSA data architecture

Compartimos follows a service-oriented approach, similar to the OGSA data architecture. In Compartimos OGSA data architecture services are specialized to make provision for address data in an SDI environment. Table 1 provides a summary overview of the services in the OGSA data architecture and their counterparts in Compartimos.

While the replica and transfer services are more generic in nature and adopted in Compartimos from the OGSA data architecture with few or no modifications, other Compartimos services are tailored specifically for address data in an SDI environment. Some aspects of the OGSA data architecture, such as policies, storage management and caching, are excluded from Compartimos because they can be used generically for any kind of data and do not have to be tailored specifically for address data.

Table 1 Services in the OGSA data architecture and related services in Compartimos

OGSA data architecture	Compartimos
Data Transfer	TransferService
Data Access	AddressDataAccessService
Storage Management	Not included in Compartimos[a]
Cache Services	Not included in Compartimos[a]
Data Replication	ReplicaService
Data Federation	VirtualAddressDataService for federation and consolidation; the OGSA data architecture only provides for federation
Data Catalogues and Registries	Catalogue and CatalogueService
Data source or resource	AddressDataset
Not in the OGSA data architecture	AddressService

[a] No need for address data specialization; generic grid-enabled services are sufficient.

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

Fig. 2 is a combination of the layered grid architectures described by Foster *et al.* [2] and Venugopal *et al.* [29]. The Compartimos components are added (in ***bold italics***) to show where

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

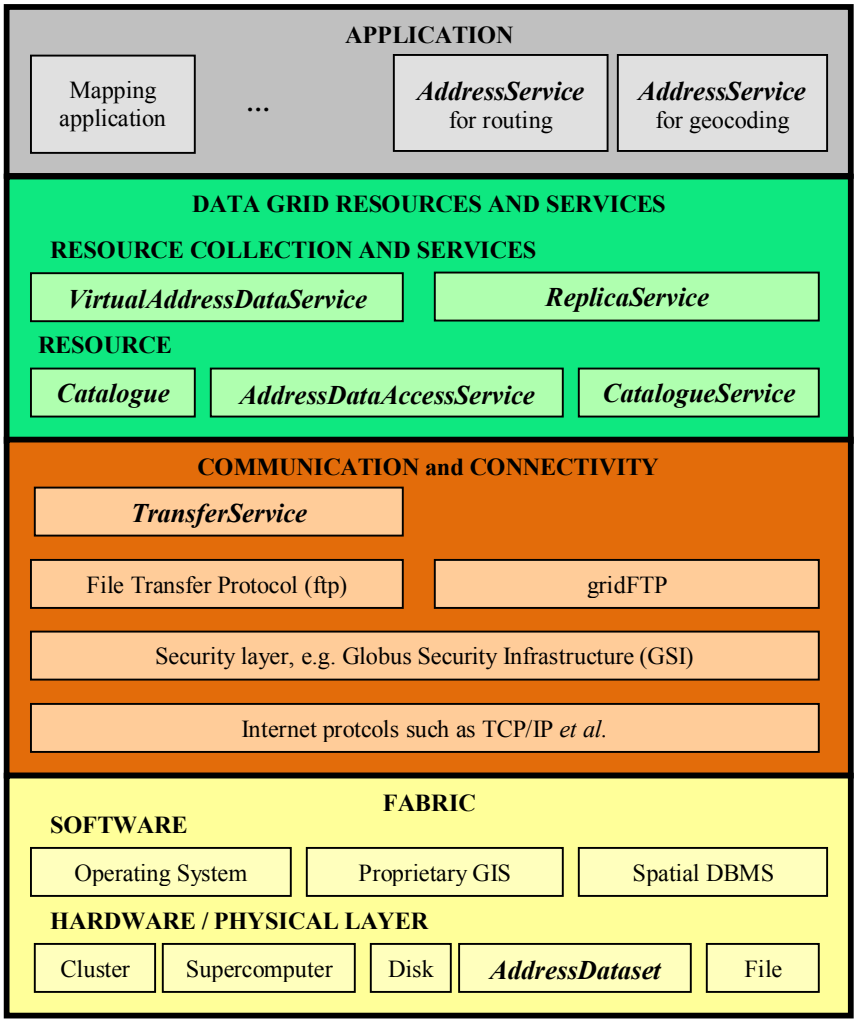

Fig. 2 The Compartimos services in the four main layers of the grid architecture

3.2 The virtual organization (VO) of an address data grid in an SDI

A VO in a grid comprises a set of individuals and/or institutions having direct access to computers, software, data, and other resources for collaborative problem solving or other

purposes. VOs are a concept that supplies a context for operation of a grid that can be used to associate users, their requests, and a set of resources. The sharing of resources in a VO is necessarily highly controlled, with resource providers and consumers defining clearly and carefully just what is shared, who is allowed to share, and the conditions under which sharing takes place.

In the data grid approach to address data in an SDI, a VO comprises a set of individuals and/or institutions having direct access to address data from various address data sources (the resources) that is presented to the users as a seamless, single virtual address dataset to which they have access through various services (the purpose). The VO member roles are described in the subsequent paragraphs.

The *address data provider* is the institution that publishes the address dataset on the data grid. This could be the custodian or owner of the data, such as a local authority, an appointed distributor of the data, such as a consultant acting on behalf of a local authority, or an independent organization collecting address data as VGI (e.g. www.openaddress.org). The address data provider produces new releases of the data and defines what data is shared, who is allowed to share the data, and the conditions under which the sharing takes place (in agreement with the owners of the data, of course).

Table 2 Member roles in a VO

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

The *address data host* is the institution that provides the required resources to host the dataset on the data grid. For this, it has to provide an implementation of the uniform interface to the underlying address dataset, as well as a hosting environment for the interface and the data itself. The data host could be the same institution as the data provider, or it could be a third party, such as an Internet service provider (ISP) or a cloud provider.

The *node host* is the institution that provides the resources to host a node in the data grid. In data grid literature the node is sometimes referred to as a point of presence (e.g. in the GEON

grid www.geongrid.org). The node hosts the catalogue and virtual address data services together with optional services for replication, transfer, etc. There are different levels of nodes depending on whether the node hosts the optional services and/or provides additional storage space for uploading address data to the grid.

The *address-related service provider* is any third party providing address-related services, such as routing or geocoding, on top of the single virtual address dataset. By definition, the service provider is also an address data consumer.

The *address data consumer* is any user (an individual user, an institution or an application) that requests data, whether for mapping, address capturing, routing or otherwise, from the address data grid. The *address-related service consumer* is any user (an individual user, an institution or an application) that consumes an address-related service such as a routing service provided on top of the data grid. A VO member could be both a data consumer as well as a service consumer.

In its simplest form the VO has members that are data providers, data hosts, node hosts and external data consumers. An institution can adopt more than one role in the VO. For example, a medium-sized local authority might be a data provider, data host and node host. Table 2 provides examples of some of the combinations of roles. A VO can be short-lived, for example, for the duration of a specific disaster relief operation; or long-term, for example, for the verification of residential addresses of new customers when applying for a financial account.

3.3 The Compartimos components

Table 3 provides an overview of the Compartimos components while Fig. 3 shows how the Compartimos components interact with each other in the address data grid. The *Consumer*, *Address data provider* and *Address service provider* components are external to Compartimos and are therefore shown in a different color. Fig. 4 shows how the Compartimos components discussed in this section relate to the VO member roles discussed in the previous section.

Table 3 Overview of the Compartimos components

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

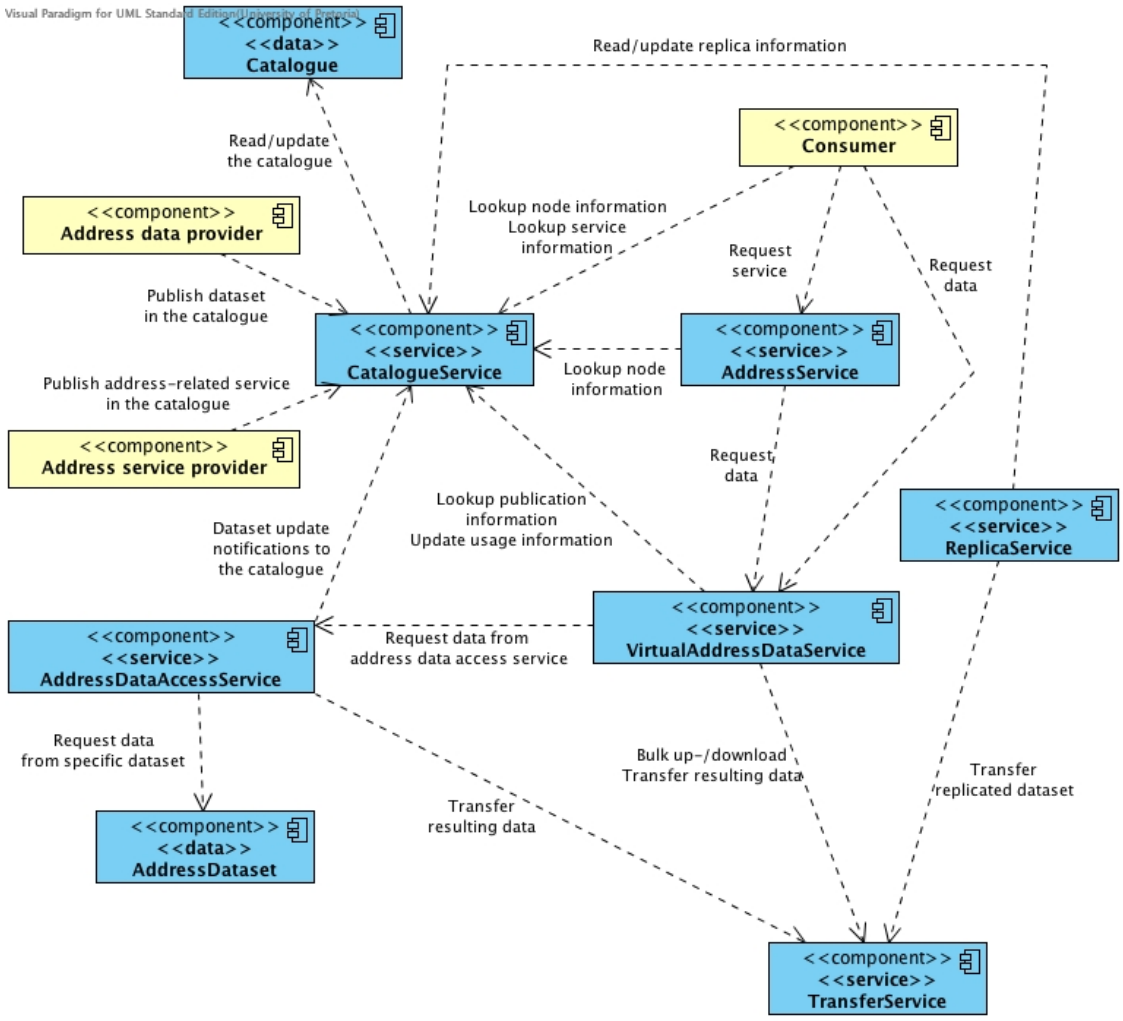

Fig. 3 Compartimos component interaction

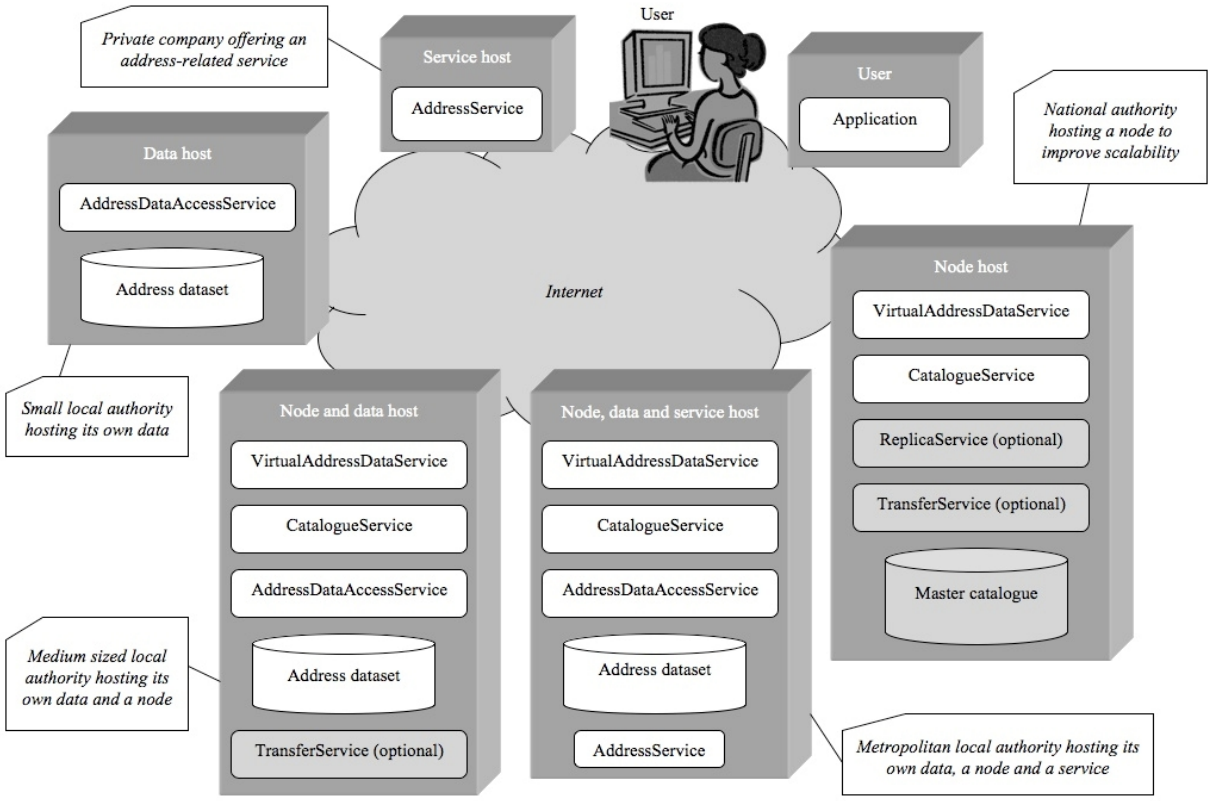

Fig. 4 VO members and Compartimos components

3.3.1 The catalogue

The Compartimos catalogue contains the metadata that is required for the operation of the address data grid. Fig. 5 shows the elements of the catalogue.

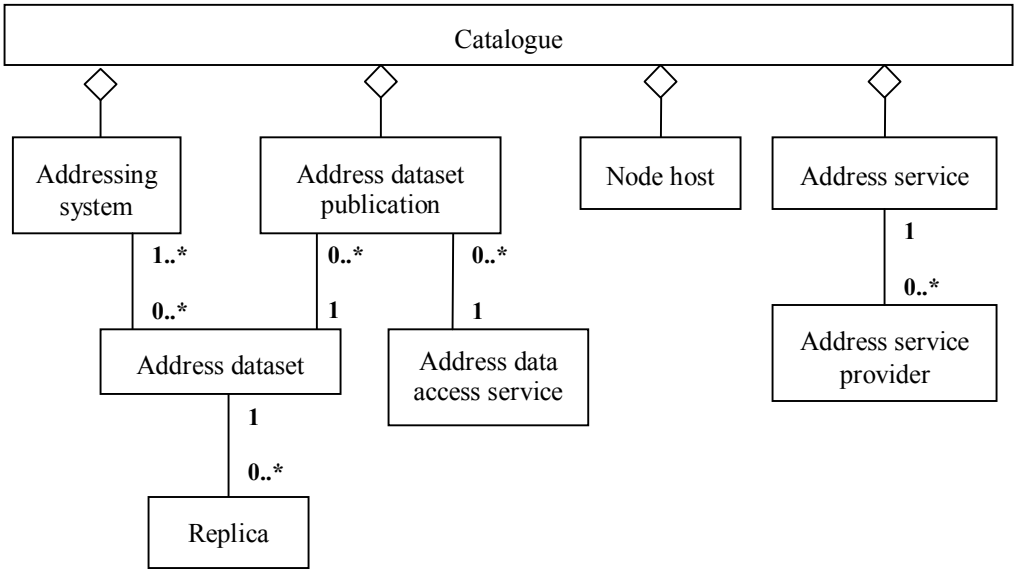

Fig. 5 The address data catalogue

The addressing systems describe the types of addresses that are contained in an address dataset, e.g. street and/or intersection address types. Details of the specialized address data model for the catalogue can be found in Coetzee [30]. A dataset is published on the address data grid by associating it with an *AddressDataAccessService*. Information about where and how a dataset is replicated is also stored in the catalogue. Address service providers provide address-related services, such as geocoding or mapping, that operate on the single virtual address dataset. The node host provides the resources to host some or all of the catalogue, replica, transfer and virtual address data service. Any interaction with the catalogue takes place through the *CatalogueService*.

The size of the catalogue is determined by the size of the catalogue's collections. Based on the number of countries in the world, in an international address data grid these numbers are relatively small in respect of what relational DBMS, object-oriented DBMS and XML databases are able to cope with, and there is no need to make special provision for huge volumes of data. The data model for the catalogue is sufficiently simple to allow representation in a relational data model. However, it is important that the storage mechanism for the catalogue is platform independent so that it can be easily replicated and XML is therefore attractive.

3.3.2 The catalogue service

The main purpose of the *CatalogueService* is to provide read and write access to the information that is stored in the Compartimos catalogue. Similar to the OGSA data architecture, the Compartimos catalogue service provides *Publish* (add an entry), *Update* (modify an existing entry), and *Find* (apply query and return matching entries) services. Because Compartimos applies to a very specific kind of data, the *Augment* (add additional properties for an entry created by someone else), *AddClassification* (add classification scheme) and *Classify* (classify an entry) services from the OGSA data architecture are not required and therefore not included. However, if Compartimos is revised to include any kind of spatial data (cadastral land parcels, points of interest, transport network, etc.), these three services will be relevant.

3.3.3 The address data access service

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

The OGSA data architecture proposes three generic data access operations for structured data: *Create*, *ExecuteQuery* and *BulkLoad*. The *Create* operation creates an association between a data service and an underlying data resource, which may be created and populated as a result of this operation.

RegisterDataPublication operation of the *CatalogueService* associates a dataset with an *AddressDataAccessService*. The Compartimos model provides for a one-to-many relationship between a dataset and an access service, allowing more than one access service to be associated with the same dataset.

In Compartimos the *ReplicaService* uses the *CreateDataset* operation of the *AddressDataAccessService* to create a replica. Once this replica of an original dataset has been created and populated, its information is added to the catalogue, and it can be used in subsequent data queries. Thus, in Compartimos the physical creation of the dataset is separated from adding the association between an address dataset and an address data access service to the catalogue. This separation is reflected in the *1*..* relationships between an address dataset publication and its associated dataset and address data access service in Fig. 5.

3.3.4 The virtual address data service

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

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

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

3.3.5 The address dataset

The Compartimos *AddressDataset* component refers to any address dataset that is published on the address data grid. In OGSA data architecture terminology this is the data source or data resource. The *RegisterDataPublication* operation of the catalogue service associates an address data access service with a particular address dataset, and from then on the *AddressDataset* is available for inclusion in address data queries and requests on the grid. The particulars of the underlying dataset, such as the format, data model, etc., influence the performance of data access.

3.3.6 The address-related service

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

3.3.7 The replica service

The *ReplicaService* is responsible for replicating address datasets for fault tolerance, faster access and for scalability reasons. Replicas of datasets are stored on additional storage that is provided at the different node hosts. A node opts to allow replication or not. Datasets are either replicated as a whole, or parts thereof. There are different ways of splitting up a dataset for replication, for example, by selecting a geographic region of the dataset, by selecting specific address types, or by selecting addresses based on their creation date. An alternative way of splitting up an address dataset is to replicate the values of higher-level location types, thus providing an index into the dataset.

The *ReplicaService* is responsible for creating, deleting, validating, modifying the contents, and synchronizing the replicas of a dataset in close coordination with the *CatalogueService*: the *ReplicaService* updates the *CatalogueService* with replica information as necessary. A dataset is replicated only if its data provider allows this by setting the appropriate attributes upon registration of the dataset in the catalogue, and the security policies of the original dataset have to be maintained by the replicas.

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

3.3.8 The transfer service

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

This service does not require customization or specialization for address data, and in Compartimos mostly the same operations as in the OGSA data architecture are included: *SetupTransfer, PauseTransfer, ResumeTransfer* and *StopTransfer*. The *CreateTransfer* service in the OGSA data architecture has is called *StartTransfer* in Compartimos, and a *GetTransferState* operation, with which the state of the transfer can be monitored, as recommended by the OGSA data architecture, is included. Similar to the *ReplicaService*, the *TransferService* is isolated as a component on its own, both conceptually as well as on implementation level, allowing the address data grid to employ different transfer services over a period of time.

3.4 Security

The OGSA data architecture includes a section on security that describes issues that are important in a data grid. Specific security-related services are not included in the OGSA data architecture (nor Compartimos) but it is recommended that all services should:

- advertise the degree to which they adhere to security requirements;
- accept security related information in their interfaces; and
- pass security related information, such as security credentials in all service requests from this service.

The above recommendations apply to Compartimos. Similarly, to ensure data privacy, the following issues need to be addressed by all services in Compartimos, as recommended in the OGSA data architecture:

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

The Compartimos address data model deliberately excludes any information about the person(s) or business residing at an address, to protect their privacy.

4 Discussion

The scenario described earlier in this paper relates to the grid definition provided in the introduction. The *non-trivial qualities of a service* are represented by the geocoding and mapping of locations of damage and distress reports. Another example of such a service relevant to the scenario would be the verification of insurance claims. The address datasets at the various cities, along with any VGI data sources at independent organizations, are the *resources that are not under centralized control* and both the datasets' heterogeneity, as well as the different computing environments at the various locations, calls for *standard, open protocols and interfaces*. Apart from decentralization, the grid also allows resources to come and go, which is important in today's more dynamic SDI.

VO member roles identified in Compartimos can also be related to the scenario. The cities and independent organizations (with VGI datasets) are *address data providers* and *address data hosts*. If the address data is downloaded to the ERC (which is suggested as an alternative strategy), the ERC also becomes a data host. The ERC was described primarily as a *node host*, but as suggested, with grid-enabled geocoding software, the cities could also become node hosts, thereby improving the scalability of the data grid. The ERC's geocoding software acts as *address data consumer* and *address-related service provider* (i.e. the geocoding service). The ERC personnel who do something with the locations of distress reports are the *address-related service consumers*.

Through the development of the Compartimos reference model, we have been able to identify the essential components and how they have to be specialized from their OGSA counterparts for an address data grid in an SDI. The Compartimos catalogue requires considerable specialization for addressing systems and address data. For example, in contrast to the OGSA data architecture, Compartimos allows more than one service to be associated with a single address dataset. Also, incoming address data requests indicate whether it is necessary to force synchronization before returning results. The virtual address data service provides address-related intelligence (e.g. handling incomplete addresses). For replication, address-specific specialization is required when deciding on which parts of an address dataset to replicate (e.g. a geographic region or higher-level location types). Also, data privacy has to be taken into account when replicating address data. In Compartimos the essential components for an address data grid have been identified but there is room for future work on the details of each of these components.

Another aspect of specialization that is worthwhile mentioning in an SDI context, is the question of trust: which address data sources can the data grid trust to be accurate? In many countries a residential address is a prerequisite for opening a financial account. If the address data grid is used for residential address verification, it is imperative that it is verified against legally valid addresses only. This can be achieved by making use of the metadata associated with an address to include only address data from custodians in the address verification. In countries, such as South Africa, where custodians for address data have not been assigned, this does not

work and one has to explore other mechanisms, such as calculating a confidence level for the address based on, for example, its occurrence in or omission from a number of address datasets.

The question of trust is even more relevant with VGI in the address data grid. In a Web 2.0 world, where the citizens become the sources for data, this assumption does not hold anymore. Citizens, living at an address, are the best available source to verify an address, but the question is whether they can be trusted to provide accurate data. Goodchild [31] and Craglia *et al.* [28] also raise this question and future work should investigate how such a 'wikification' of address data can be integrated into Compartimos. For example, the group responsible for the research project described in this paper has initiated work in this line, currently investigating the role of volunteered geographic information in an SDI [32].

Compartimos has been implemented as a proof of concept in a controlled environment at the University of Pretoria. 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, such as, replicating parts of an address dataset and handling of address-specific security issues. Compartimos is an *abstract* representation of the essential coponents of an address data grid in an SDI, implying that details are reduced when identifying the essential components through a process of abstraction. However, these details, such as potential overlap between replicated datasets, performance and reliability issues of distributed hosts, etc. have to be investigated and addressed in an actual implementation.

One option for local authorities is to invest servers and bandwidth, another is to buy scalable computing power and data storage in a cloud without having to support a local IT infrastructure. However, while there is more than one way of implementing Compartimos, the essential components stay the same. For example, the Compartimos 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. These different node host configurations are flexible enough to be deploye on a local infrastrcuture or in a cloud.

As part of the research project, we also analyzed existing technologies, such as the Globus Toolkit, ISO 19100 standards and OGC web service implementation specifications, for their potential use in Compartimos [33]. This analysis showed that there is a need for collaboration between grid and geospatial communities to ensure harmonization between the respective standards and tools. Current work in the joint SDI and data grid domain mainly explores grid-enabling the OGC WPS, but from the Compartimos research it is evident that grid-enabling spatial data integration in an SDI environment should also be explored. This implies that other web services specified by OGC, such as the WFS, should also be grid-enabled. Such a grid-enabled WFS could then be used as a base class for an address data access service. In the bigger picture, the ISO 19100 series of standards together with OGC implementation specifications have been implemented in a number of SDIs [34] and all of these have to be grid-enabled in order to grid-enable the SDIs. Aloisio *et al.* [20] and Yue *et al.* [24] write about such efforts, but more implementations are required to better understand the challenges under different circumstances. Such implementations would also promote the development of tools to streamline the grid-enablement.

Compartimos was developed for address data in an SDI and future research should expand Compartimos to other types of spatial data. Incorporating recent research findings on ontologies for geospatial data would be relevant [35 and 36]. A reference model for data grids that caters for all kinds of geographic information could be seen as the first step along the path of standardizing geospatial data grids. Also, research is required to better understand the requirements for grid-enabling SDIs in terms of non-technical aspects, such as policies, legislation, agreements, human and economic resources, and organizational aspects.

Finally, this research project was initiated before the current hype of 'cloud computing'. Clouds, such as those provided by Amazon, IBM, Google, Microsoft and the like, also stand in line as the enabling platform in an SDI. Grids arose to address large-scale computation problems using a network of resource-sharing commodity machines, resulting in a focus on integration of existing resources. In contrast, clouds are developed to address Internet-scale computing problems with different assumptions as for the grid: clouds refer to large pools of computing and/or storage resources that can be accessed via standard protocols and can be built on existing protocols. In addition, the business model in the cloud is pay-per-use while in grids it is mostly project-oriented [37]. Thus, clouds have the potential to address a different requirement of address data in an SDI, which needs to be researched.

5 Conclusion

In this paper an emergency response scenario was presented to illustrate how the data grid approach can solve the problem of decentralized address data in a dynamic SDI. This approach is both a novel application for data grids as well as a novel technology in SDI environments, and thus improves the understanding of the requirements and issues related to applying the data grid approach in an SDI. The paper further presented Compartimos, a reference model for an address data grid in an SDI based on the OGSA data architecture. Compartimos identifies the essential components and their capabilities that are required for a decentralized address data grid in a dynamic bottom-up SDI. Address- and SDI-specific capabilities have been designated to the components, thus indicating where specialization for address data in an SDI is required. In conclusion, additional research that is required in the area of address data grids for SDIs is discussed. This includes, but is not limited to grid-enabling spatial data integration in an SDI; the issue of trust and data sources in a grid; understanding the requirements for grid-enabling SDIs in terms of non-technical aspects, such as policies, legislation, agreements, human and economic resources, and organizational aspects; and the value of clouds for address data in an SDI.

Acknowledgements This research was made possible via the support of AfriGIS (www.afrigis.co.za) and the Technology and Human Resources for Industry Programme (THRIP) managed by the South African National Research Foundation (NRF) and financed by the South African Department of Trade and Industry.

References

1. Open Grid Forum (OGF) (2007) *Open Grid Services Architecture Glossary of Terms Version 1.6*, Editor: J. Treadwell. http://www.ogf.org/gf/docs/?final. Accessed 23 April 2010.
2. Foster I, Kesselman C and Tuecke S (2001) The Anatomy of the grid – enabling scalable virtual organizations. *International Journal of High Performance Computing Applications*. **15**(3), pp200-222.
3. Foster I, Kesselman C, Nick JM and Tuecke S (2002) *The Physiology of the grid: An Open Grid Services Architecture for Distributed Systems Integration*. http://forge.gridforum.org/sf/go/doc13483?nav=1. Accessed 23 April 2010.
4. Open Grid Forum (OGF) (2006) *The Open Grid Services Architecture*, Version 1.5. Editors: I. Foster, H. Kishimoto and A. Savva. http://www.ogf.org/gf/docs/?final. Accessed 23 April 2010.
5. Open Grid Forum (OGF) (2007) *OGSA Data Architecture*. Editors: D Berry, A Luniewski, M Antonioletti. http://www.ogf.org/gf/docs/?final. Accessed 23 April 2010.
6. Crompvoets J, Bregt A, Rajabifard A, Williamson I (2004) Assessing the worldwide developments of national spatial data clearinghouses. *International Journal of Geographical Information Science*, October-November 2004; **18**(7), pp665-689.
7. Masser I, Rajabifard A, Williamson I (2008). Spatially enabling governments through SDI implementation. *International Journal of Geographic Information Science*, January 2008; **22**(1), pp 5-20.
8. Levoleger K and Corbin C (Ed.) (2005) *Survey of European National Addressing as of May 2005, European Umbrella Organisation for Geographic Information (EUROGI)*. http://www.eurogi.org/POOLED/articles/bf_docart/view.asp?Q=bf_docart_184387. Accessed 23 April 2010.
9. Coetzee S, Cooper AK, Lind M, McCart Wells M, Yurman SW, Wells E, Griffiths N and Nicholson MJL (2008). Towards an international address standard. *GSDI-10 Conference, Trinidad and Tobago*, 25 – 29 February 2008.
10. Paull D (2003). *A Geocoded National Address File for Australia: The G-NAF What, Why, Who and When*. PSMA Australia Limited. http://www.psma.com.au/products/gnaf.cfm. Accessed 23 April 2010.
11. Fahey D and Finch F (2006) GeoDirectory Technical Guide. *An Post GeoDirectory Limited*. http://www.geodirectory.ie/Downloads.aspx. Accessed 23 April 2010.
12. Nicholson M (2007). The address: improving public service delivery. *45th Annual URISA Conference*, Washington DC, USA, 20-23 August 2007.
13. Foster I (2002) What is the Grid? A three point checklist. *GRIDToday*, 22 July 2002, **1**(6).
14. Coetzee S and Bishop J (2009). Address databases for national SDI: Comparing the novel data grid approach to data harvesting and federated databases. *International Journal of Geographic Information Science*, September 2009; **23**(9), pp1179-1209.
15. Hua L, De-ren L, Xin-yan Z (2005) Large volume spatial data management based on grid computing. *Proceedings of the 2005 IEEE International Geoscience and Remote Sensing Symposium (IGARSS '05)*, 25-29 July 2005.
16. Aloisio G, Cafaro M, Fiore S, Wuarta G (2005) A grid-based architecture for earth observation data access. *2005 ACM Symposium on Applied Computing*, Santa Fe, New Mexico, USA, 13-17 March 2005.
17. Allen G, Bodgen P, Creager G, Dekate C, Jesch C, Kaiser H, McLaren J, Perrie W, Stone GW and Zhang X (2008) Towards an integrated GIS-based coastal forecast workflow. *Concurrency and Computation: Practice and Experience*, 2008; **20**(14), pp1617-1635.

18. Aydin G, Sayar A, Gadgil H, Aktas M, Fox GC, Ko S, Bulut H and Pierce ME (2008) Buliding and applying geographical information system grids. *Concurrency and Computation: Practice and Experience*, 2008; **20**(14), pp 1653-1695.
19. Xue Y, Wan W, Li Y, Guang J, Bai L, Wang Y, and Ai J (2008). Quantitative retrieval of geophysical parameters using satellite data. *IEEE Computer*, April 2008; **41**(4), pp33-39.
20. Aloisio G, Cafaro M, Conte D, Tiore S, Epicoco I, Marra GP, Quarta G (2005). A grid-Enabled Web Map Server. *International Conference on Information Technology: Coding and Computing (ITCC'05) - Volume I*, 2005, pp298-303.
21. Shu Y, Zhang JF, Zhou X (2006). A grid-Enabled Architecture for Geospatial Data Sharing.s *IEEE Asia-Pacific Conference on Services Computing (APSCC '06)*, December 2006, pp369 – 375.
22. Wei X, Yue P, Dadi U, Min M, Hu C, Di L (2006) Effective Acquisition of Geospatial Data Products in a Collaborative grid Environment. *IEEE International Conference on Services Computing*, SCC '06, September 2006, pp455-462.
23. Padberg A and Kiehle C (2009). Towards a Grid-Enabled SDI: Matching the Paradigms of OGC Web Services and Grid Computing. *Proceedings of the Eleventh Conference of the Global Spatial Data Infrastructure Association (GSDI-11)*, Rotterdam, Netherlands, 15-19 June 2009. http://www.gsdi.org/gsdi11/papers/pdf/174.pdf. Accessed 23 April 2010.
24. Yue P, Gong J, Di L, He L and Wei Y (2009) Integrating semantic web technologies and geospatial catalog services for geospatial information discovery and processing in cyberinfrastructure. *Geoinformatica*, doi 10.1007/s10707-009-0096-1.
25. Hui L, Jun Zhu, Gong J, Bingli X, Hua Q (2010) A grid based collaborative virtual geographic environment for the planning of silt dam systems, *International Journal of Geographical Information Science*, **24**(4), pp607-621.
26. *GIS.Science (2009). Die Zeitschrift für Geoinformatik*, ISSN 1430-3663, March 2009.
27. OGC OGF Memorandum of Understanding (2007), in *Directions Magazine*. Available at: http://apb.directionsmag.com/index.php?url=archives/3567-OGC-OGF-MOU.html&serendipity[cview]=linear Accessed 23 April 2010.
28. Craglia M, Goodchild MF, Annoni A, Camara G, Gould M, Kuhn W, Mark D, Masser I, Maguire D, Liang S and Parsons E (2008). Next-Generation Digital Earth, a position paper from the Vespucci Initiative for the advancement of Geographic Information Science. *International Journal of Spatial Data Infrastructures Research*, 2008;Volume 3, pp146-167.
29. Venugopal S, Buyya R and Ramamohanarao K (2006). A taxonomy of data grids for distributed data sharing, management and processing. *ACM Computing Surveys*, March 2006, Vol. 38, Article 3, pp.1-53.
30. Coetzee S (2009) *An analysis of a data grid approach for spatial data infrastructures*. PhD dissertation, University of Pretoria, South Africa. http://upetd.up.ac.za/thesis/available/etd-09272009-152926/. Accessed 23 April 2010.
31. Goodchild MF (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, Volume 69, pp211-221.
32. Cooper AK, Coetzee S and Kourie D (2010). *Perceptions of virtual globes, volunteered geographical information and spatial data infrastructures*, *Geomatica,* **64**(1), pp73-88.
33. Coetzee S and Bishop J (2009). An analysis of technology choices for data grids in a spatial data infrastructure. *SDI Convergence. Research, Emerging Trendss, and Critical Assessment*, Van Loenen B, Besemer JWJ and Zevenbergen JA (eds.). Nederlandse Commissie voor Geodesie (Netherlands Geodetics Commission), 2009; 48, pp107-120.

34. Aalders HJGL (2005). An introduction to spatial metadata standards in the world, in *World Spatial Metadata standards*, edited by Moellering H, Aalders HJGL and Crane A, Elsevier, Oxford, United Kingdom, 2005.
35. Alam A, Khan L and Thuraisingham B (2010) Geospatial Resource Description Framework (GRDF) and security constructs. *Computer Standards & Interfaces*, doi:10.1016/j.csi.2010.01.002.
36. Brisaboa NR, Luaces MR, Places AS and Seco D (2010) Exploiting geographic references of documents in a geographical information retrieval system using an ontology-based index, *Geoinformatica*, **14**(3), pp307–331, doi 10.1007/s10707-010-0106-3.
37. Foster I, Zhao Y Raicu I and Lu S (2010). Cloud Computing and grid Computing 360-Degree Compared. *2008 Grid Environments Workshop (GCE '08)*, Austin, Texas (USA), 12-16 November 2008, pp1-10.