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

4.1 Introduction

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

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

4.1.1 The purpose of Compartimos

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

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

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

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

Compartimos relates to three existing reference models:

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

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

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

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

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

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

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

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

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

4.1.2 Presentation of Compartimos

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

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

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

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

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

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

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

4.2 Enterprise viewpoint

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

4.2.1 Scope and purpose

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

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

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

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

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

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

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

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

4.2.2 High-level use cases

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

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

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

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

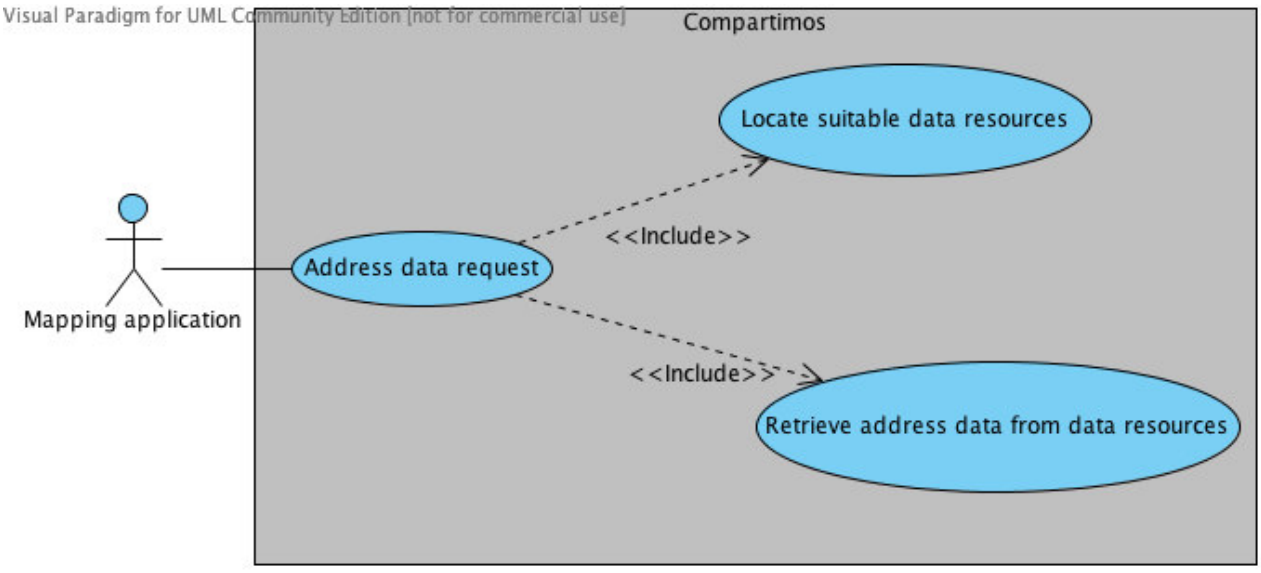

Figure 17. Simple data request (use case)

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

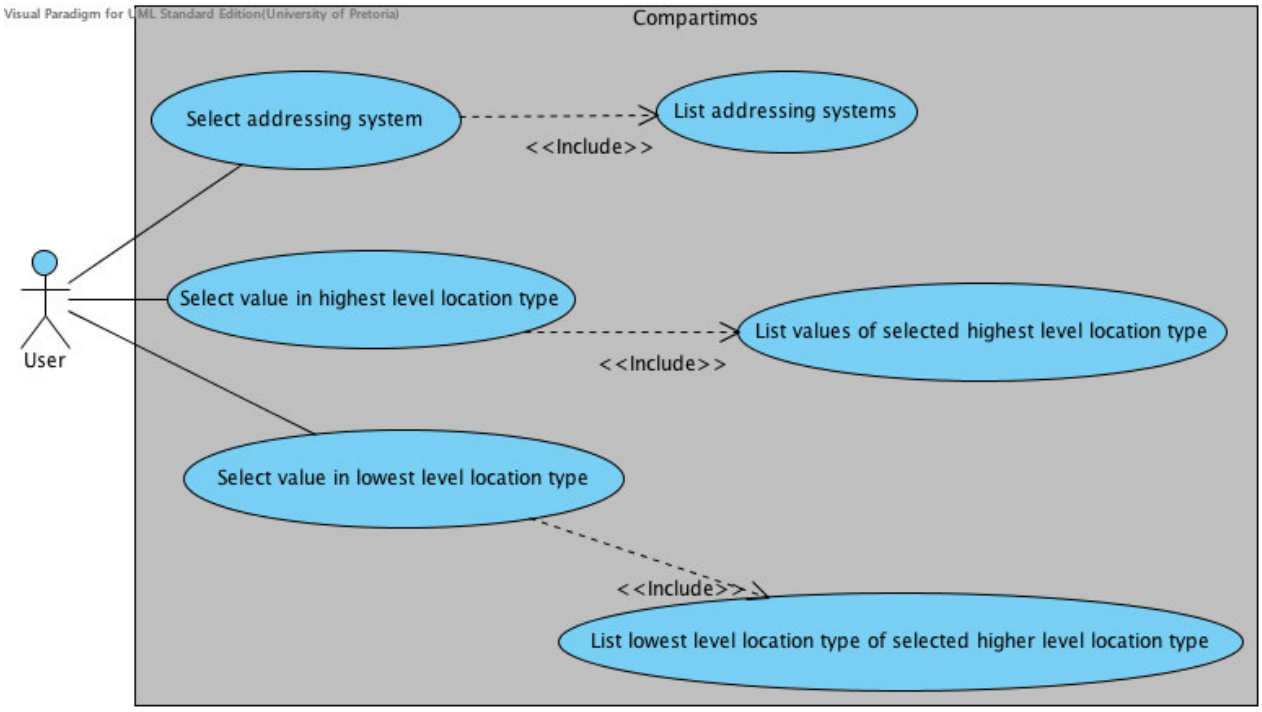

Figure 18. Iterative data request (use case)

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

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

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

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

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

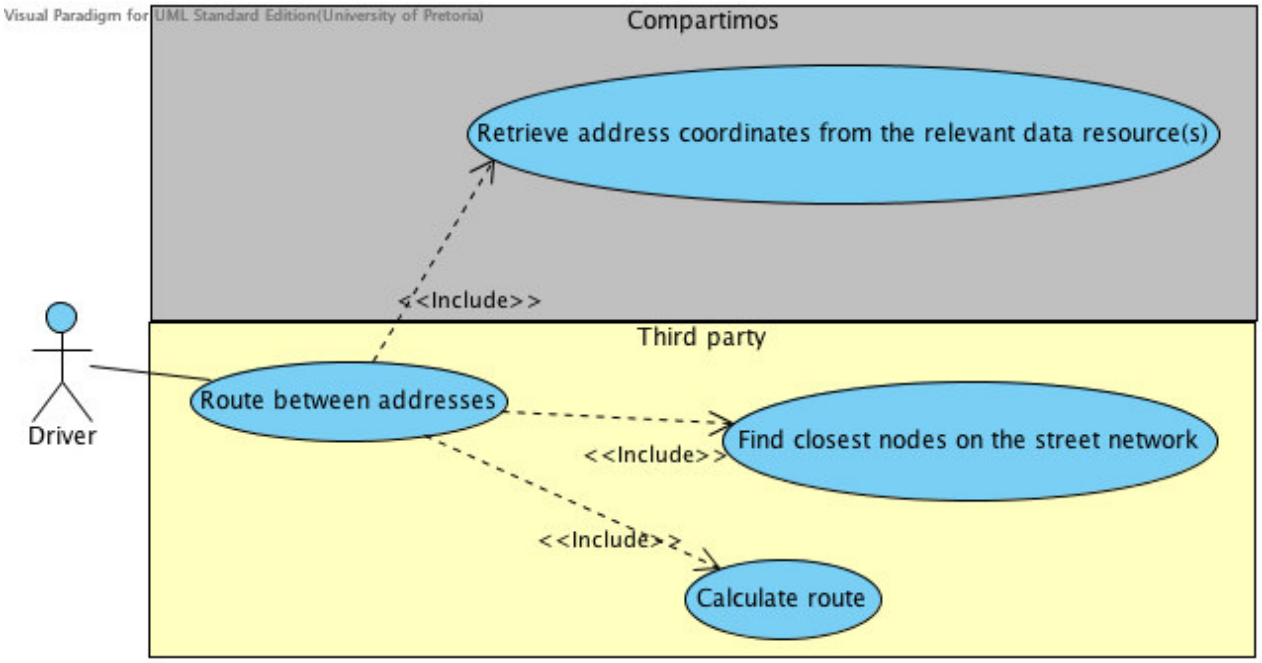

Figure 19. Service request (use case)

4.2.3 Virtual organization (VO)

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

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

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

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

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

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

Table 6. Member roles in a VO

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

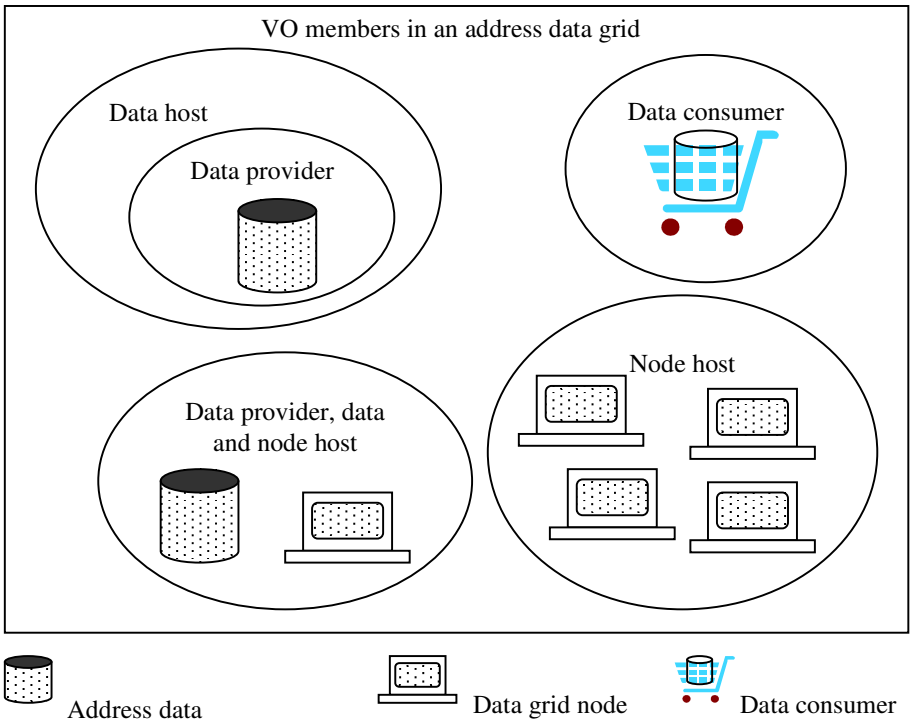

Figure 20. The VO distributed across different administrative domains (represented by the ovals)

4.2.4 SDI environment of address data

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

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

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

Figure 21. Potential for address ambiguity

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

4.3 Information viewpoint

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

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

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

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

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

4.3.1 Address data

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

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

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

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

Table 7. South African sample addresses

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

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

```
StreetAddress = StreetIdentifier, Locality

StreetIdentifier = [CompleteAddressNumber | StreetNumberRange],
                   CompleteStreetName

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

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

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

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

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

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

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

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

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

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

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

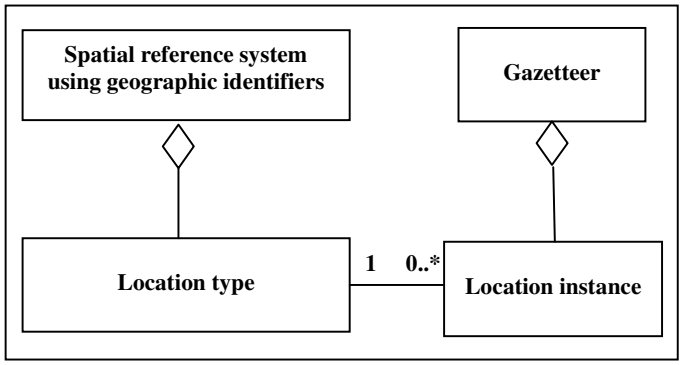

Figure 23. Spatial referencing using geographic identifiers (ISO 19112:2003)

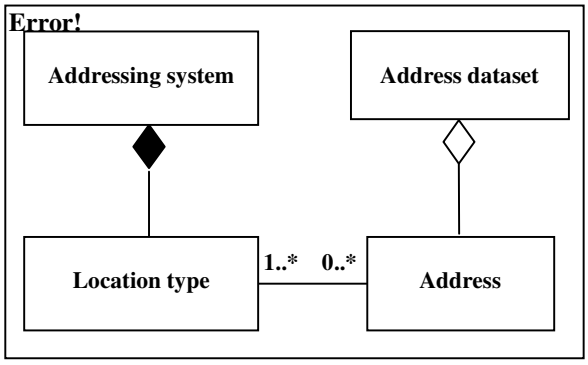

Figure 24. Spatial referencing using addresses (adapted from ISO 19112:2003)

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

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

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

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

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

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

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

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

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

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

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

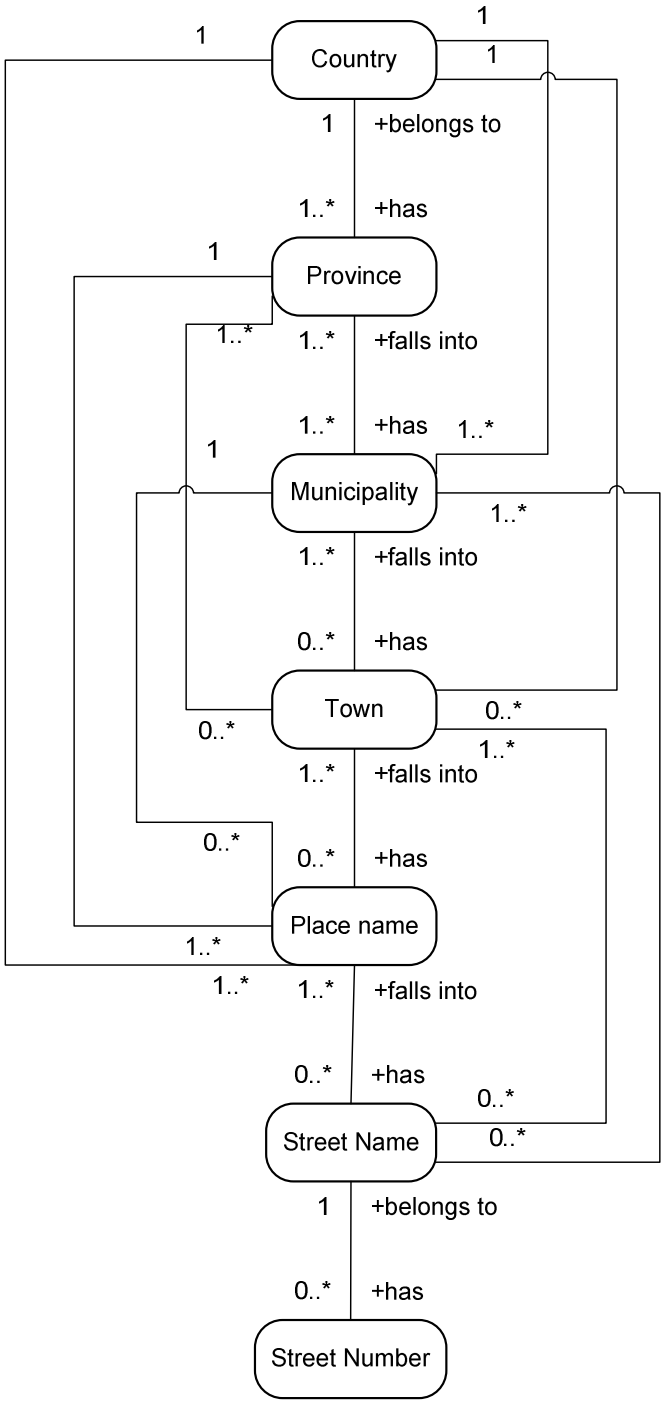

Figure 26. Relationships between the location types of the SANS 1883 street address type

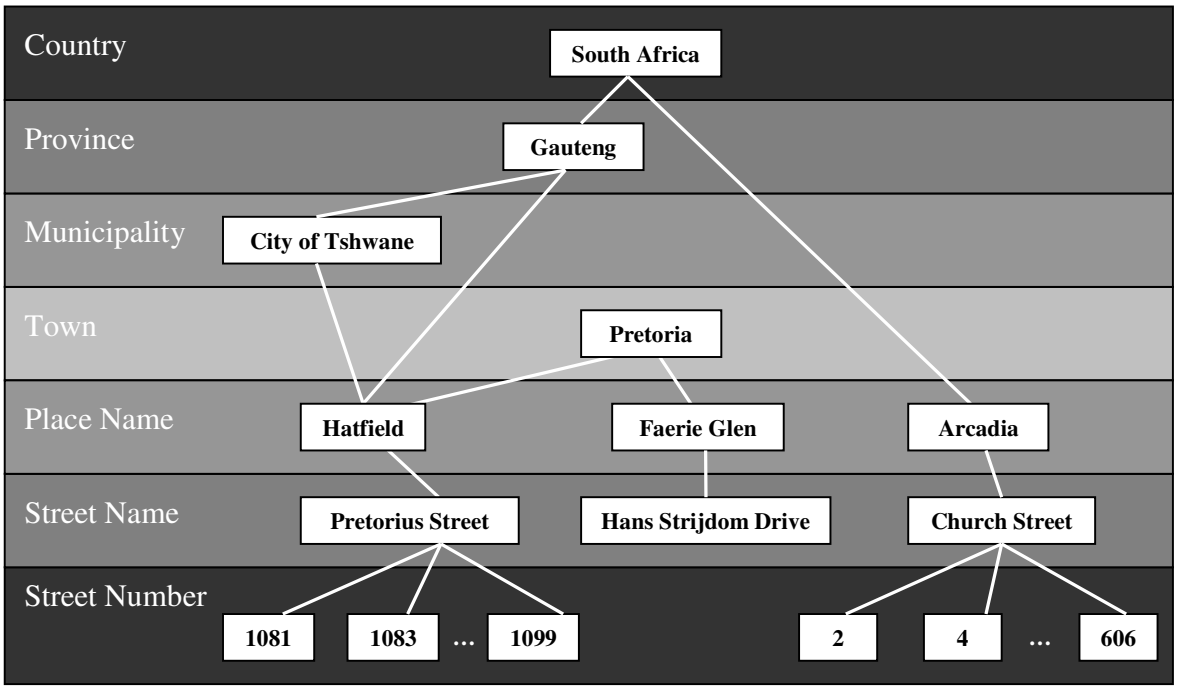

Figure 27. Instances of valid combinations of location types in the SANS 1883 street address type

4.3.2 Address data catalogue

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

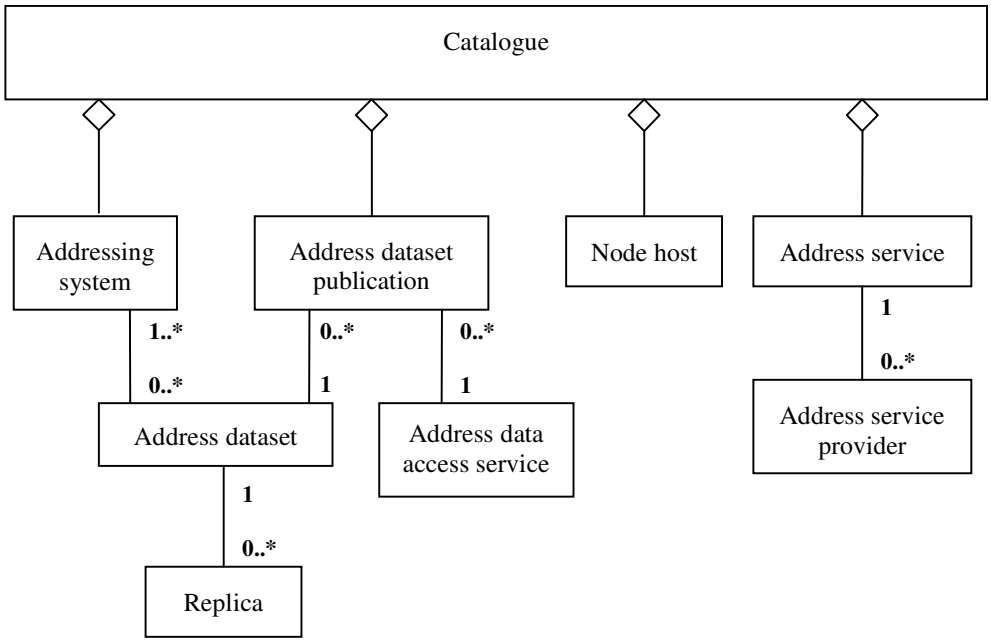

Figure 28. The address data catalogue

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

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

4.4 Computational viewpoint

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

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

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

4.4.1 Object overview

Compartimos comprises the following objects:

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

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

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

Table 11. Overview of the objects in the reference model

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

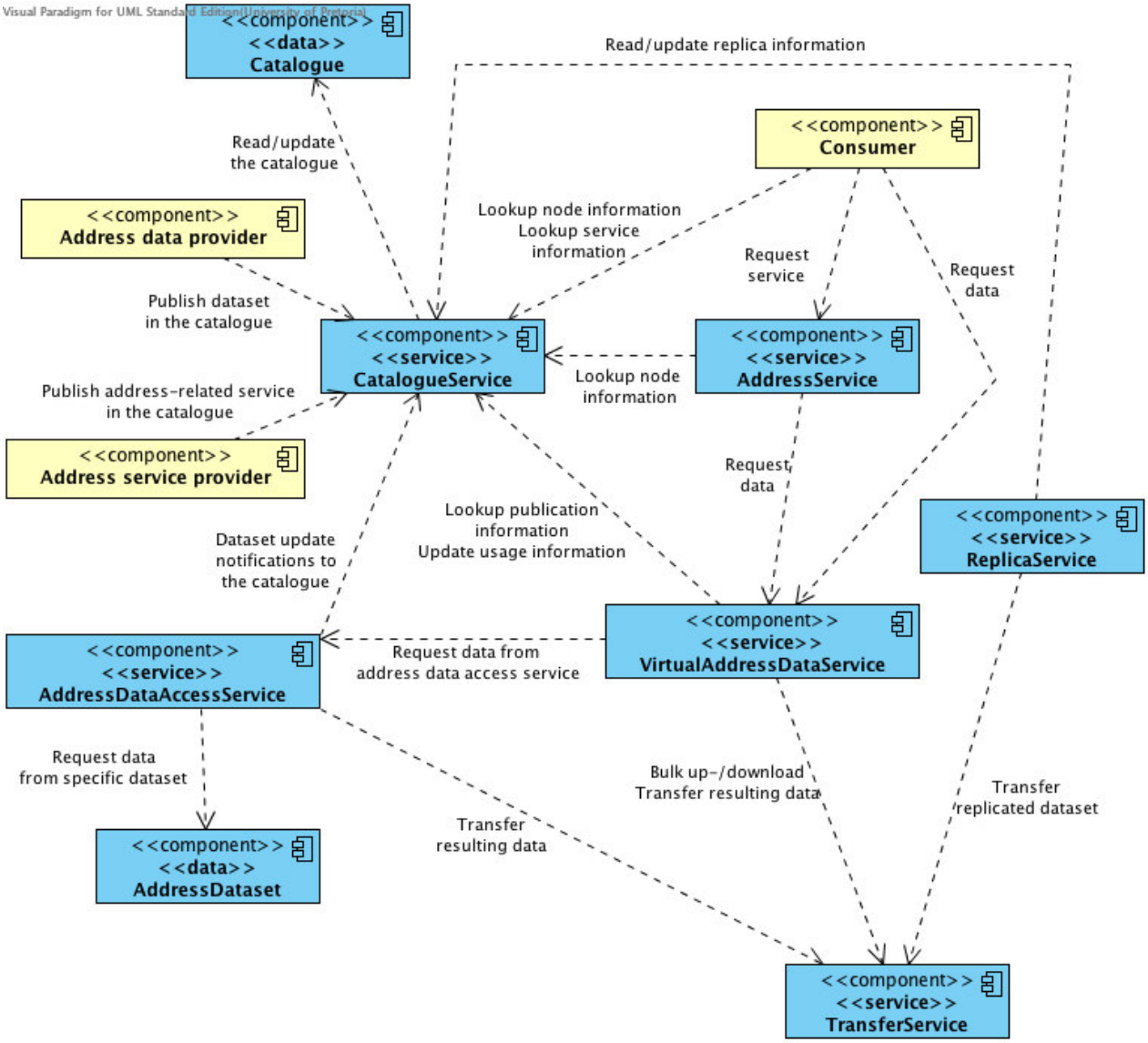

Figure 29. Object interaction in Compartimos

4.4.2 The catalogue service (CatalogueService)

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

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

4.4.3 The catalogue (Catalogue)

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

4.4.4 The replica service (ReplicaService)

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

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

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

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

4.4.5 The transfer service (TransferService)

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

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

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

4.4.6 The address data access service (AddressDataAccessService)

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

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

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

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

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

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

4.4.7 The virtual address data service (VirtualAddressDataService)

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

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

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

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

4.4.8 The address dataset (AddressDataset)

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

4.4.9 The address-related service (AddressService)

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

4.4.10 Object interaction

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

4.4.10.1 Simple data request

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

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

4.4.10.2 Iterative data request

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

4.4.10.3 Service request

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

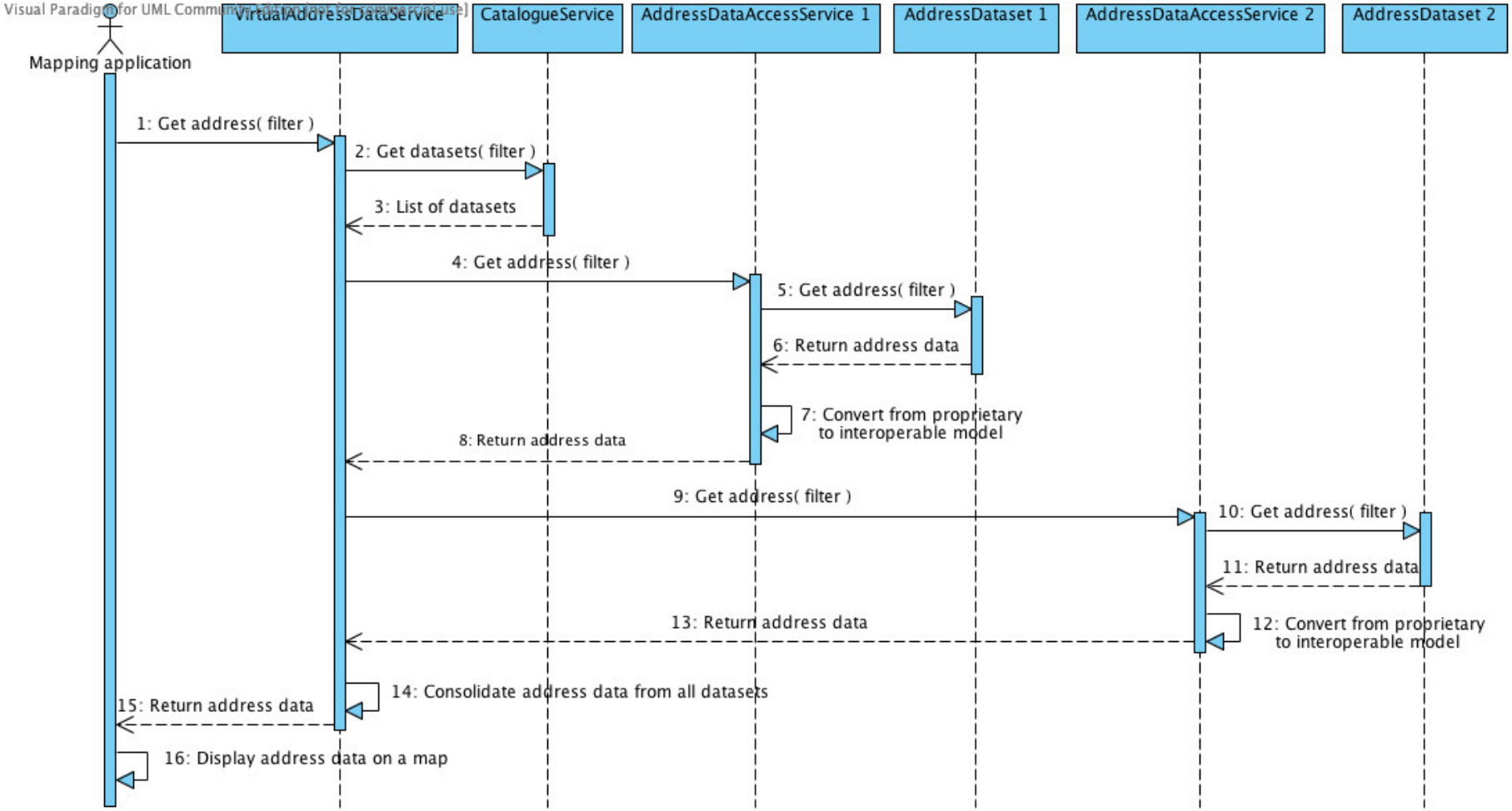

Figure 30. Simple data request (sequence diagram)

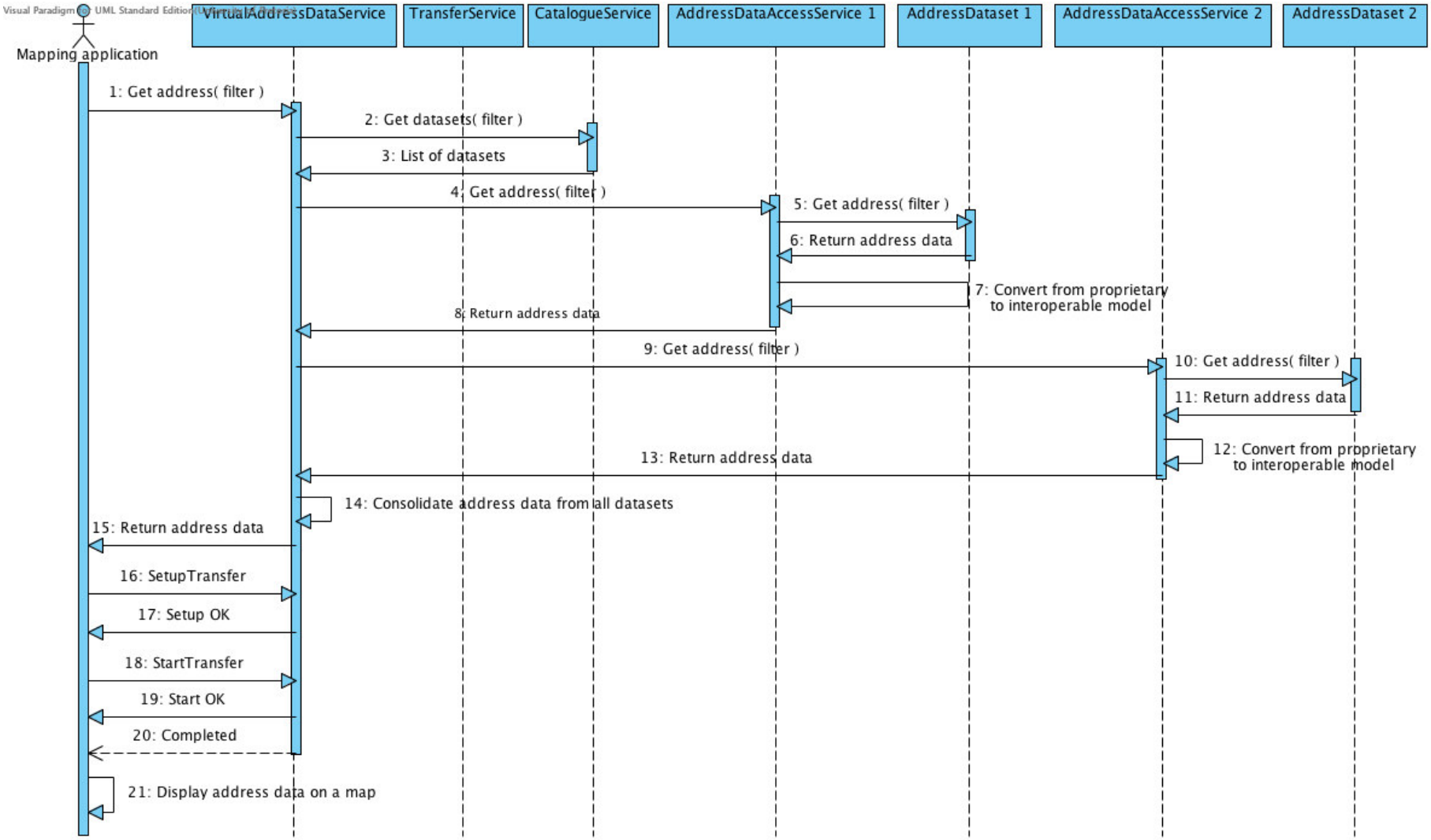

Figure 31. Simple data request involving the TransferService (sequence diagram)

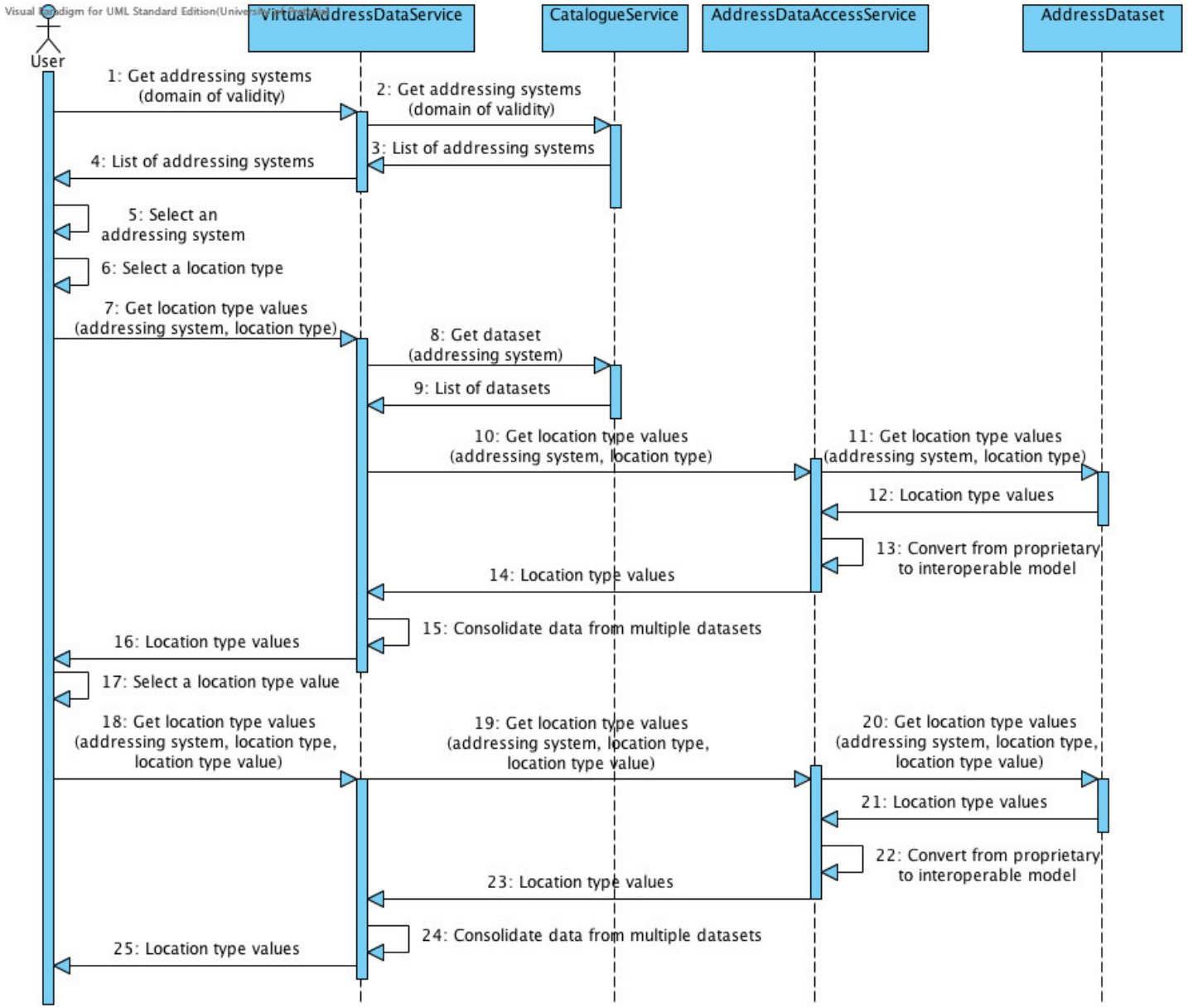

Figure 32. Iterative data request (sequence diagram)

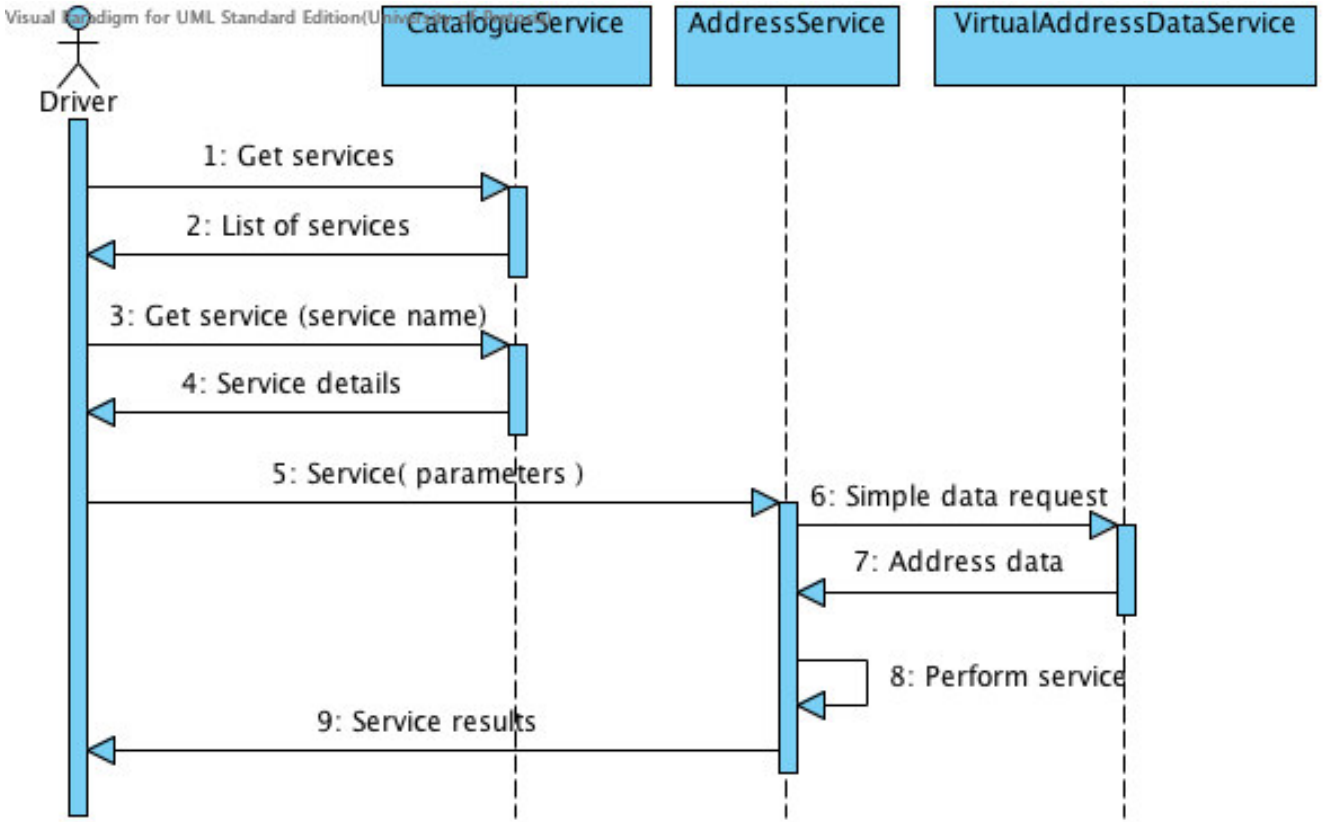

Figure 33. Service request (sequence diagram)

4.5 Engineering viewpoint

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

4.5.1 Object deployment

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

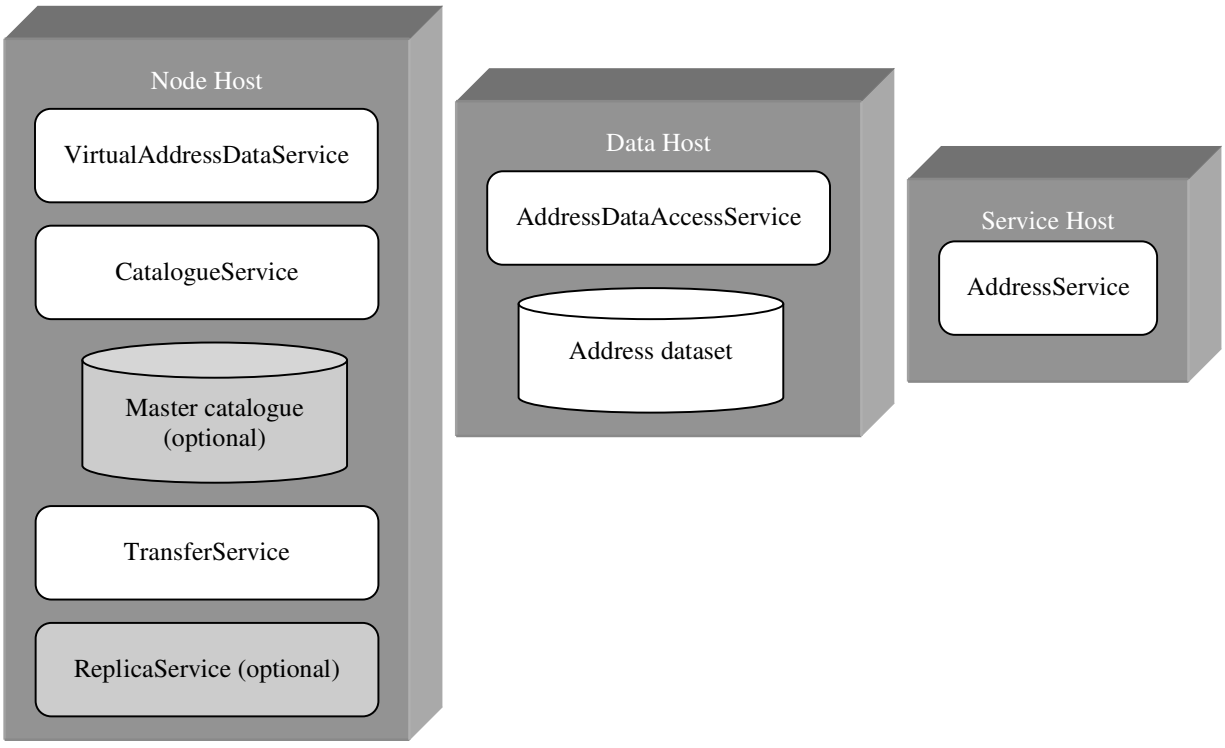

Figure 34. Three types of hosts in Compartimos

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

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

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

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

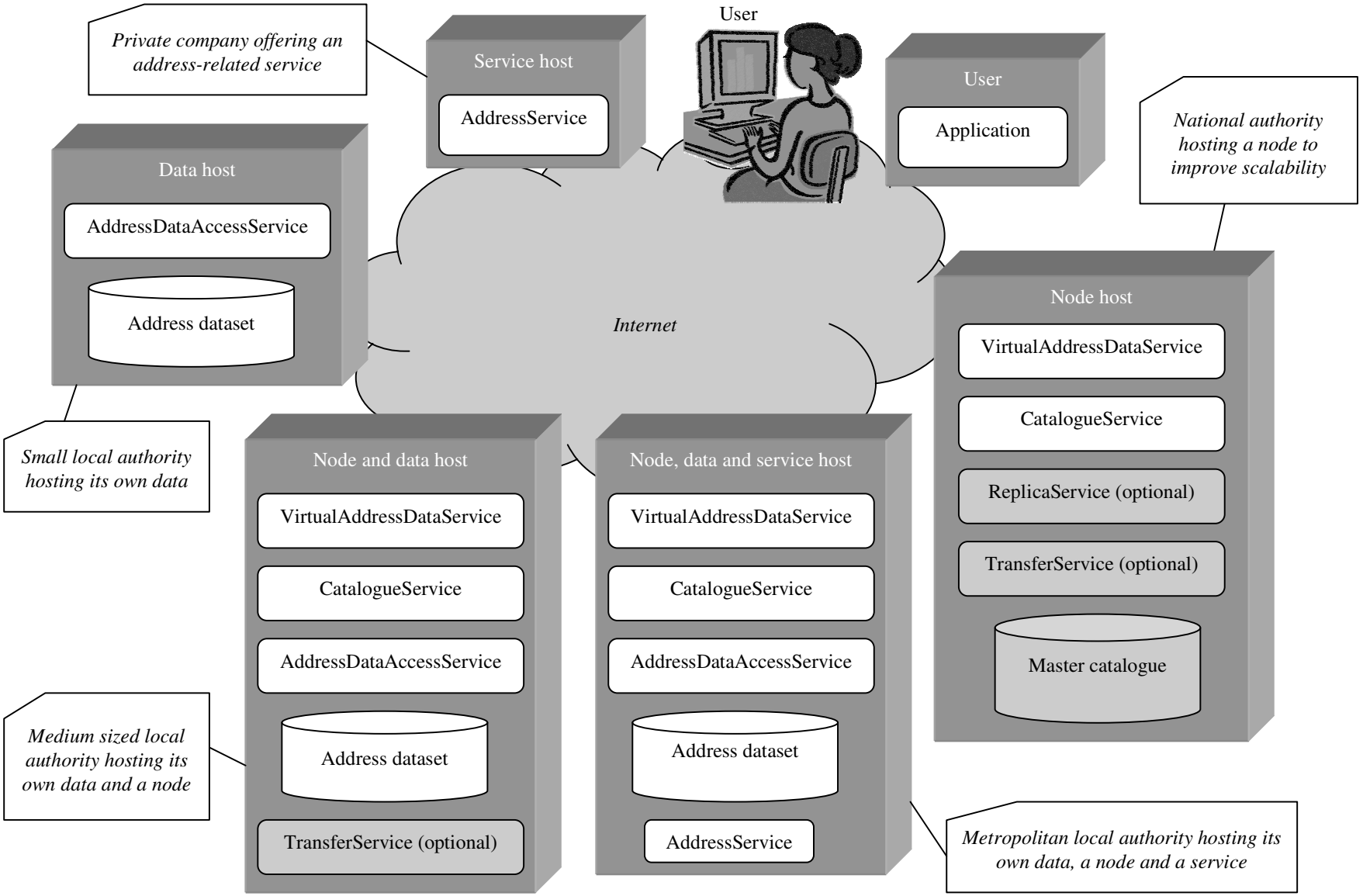

Figure 35. Deployment diagram for the address data grid with a variety of hosts

4.6 Discussion

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

4.6.1 Comparison overview to the OGSA data architecture

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

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

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

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

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

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

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

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

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

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

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

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

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

4.6.2 The layered aspect of Compartimos

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

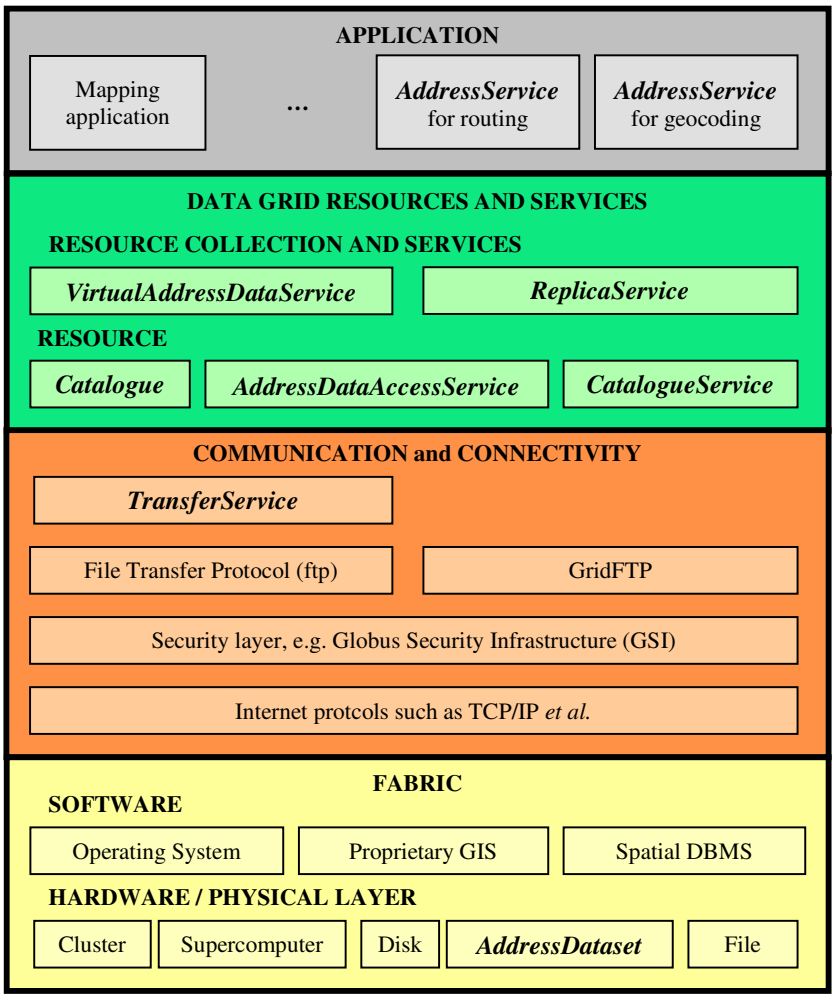

Figure 36. The Compartimos services in the four main layers of the Grid architecture

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

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

4.6.3 Service-oriented architecture of Compartimos

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

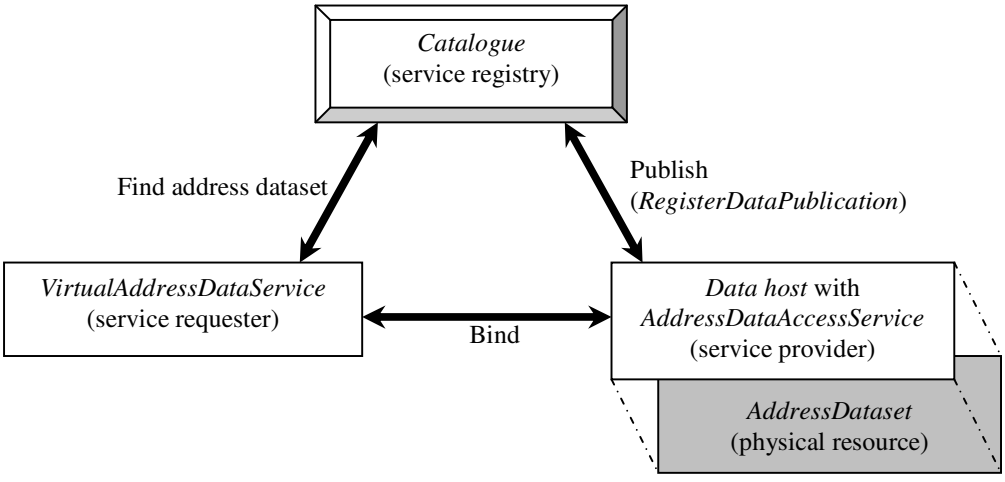

Figure 37. Service-orientation for address data access in Compartimos

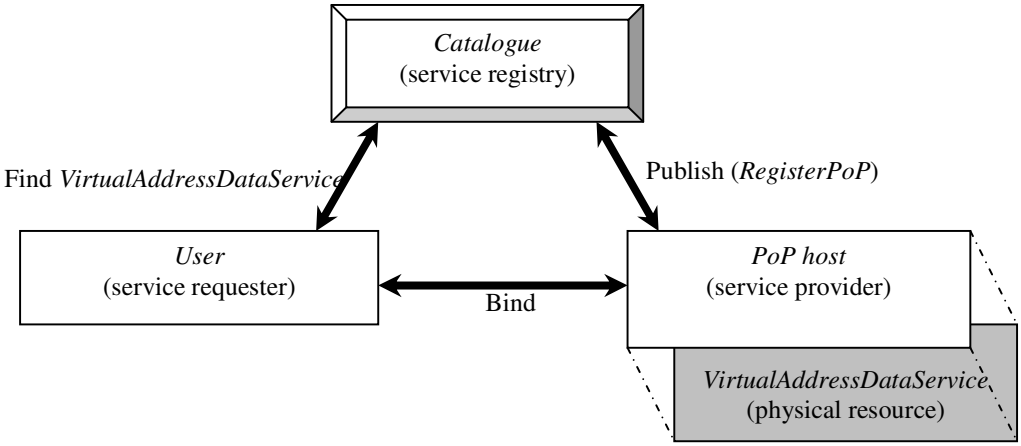

Figure 38. Service-orientation for nodes in Compartimos

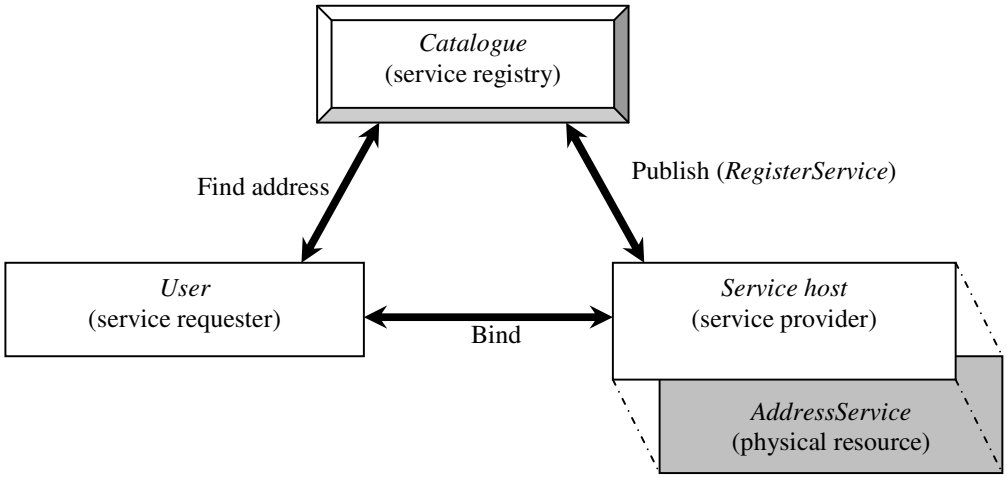

Figure 39. Service-orientation for address-related services in Compartimos