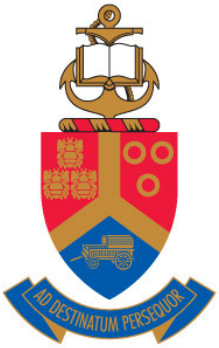

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Department Computer Science

An analysis of a data grid approach for spatial data infrastructures

by

Serena Martha Coetzee

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

Philosophiae Doctor (Computer Science)

in the

Faculty of Engineering, Built Environment and Information Technology,

University of Pretoria, Pretoria

November 2008

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

Abstract

An analysis of a data grid approach for spatial data infrastructures

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

The concept of grid computing has permeated all areas of distributed computing, changing the way in which distributed systems are designed, developed and implemented. At the same time 'geobrowsers', such as Google Earth, NASA World Wind and Virtual Earth, along with in-vehicle navigation, handheld GPS devices and maps on mobile phones, have made interactive maps and geographic information an everyday experience. Behind these maps lies a wealth of spatial data that is collated from a vast number of different sources. A spatial data infrastructure (SDI) aims to make spatial data from multiple sources available to as wide an audience as possible. Current research indicates that, due to a number of reasons, data sharing in these SDIs is still not common.

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

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

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

Preamble

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Town, South Africa.

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

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

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

Acknowledgements

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

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

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

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

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

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

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

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

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

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

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

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

Table of Contents

List of Tables

List of Figures