1 Introduction

1.1 Problem Definition

The rapid advancement in spatial data capture technologies, such as the Global Positioning System (GPS), satellite imaging and total stations, has made the capture of digital spatial data a relatively quick and easy process. This has resulted in the amount of digital spatial data in existence increasing significantly over the last decade or so. There is now a vast amount of spatial data, stored by numerous organisations at various locations across the globe. Much of these data are not being used as effectively as they should. Referring to the LANDSAT images, the Vice President of the USA said:

> In spite of the great need for the information, the vast majority of those images have never fired a single neuron in a single human brain. Instead, they are stored in electronic silos of data (Gore 1998).

Recently there has been a greater focus on how to use the spatial data that are collected and stored in the expansive "electronic silos" to their full potential (Gore 1998). Spatial Data Infrastructures (SDIs) are a key component in allowing for the better utilisation of spatial data throughout the spatial data industry. Current SDI technology allows search engines to search metadata systems within a network to discover what spatial data is in existence. It is expected that as SDIs evolve spatial data sets will be accessed "live" from their distributed locations, rather than being downloaded before being used. With this development, it is anticipated that metadata systems will evolve into metadata engines and will again be background tools that are used for querying spatial data sets that are distributed across a network.

In 1996 Buehler and McKee said:

> The dominant computing paradigm is moving away from closed systems to open systems, away from isolated systems to systems that interoperate in real time, away from tightly wrapped independent applications to application environments equipped with software components that interoperate to provide more flexible capabilities for the user.

This quote from Buehler and McKee, of the OpenGIS consortium, probably best describes the direction that GIS is heading. Users of GISs are now at the stage where they want more than just their closed systems that only have access to the spatial data that they have stored locally. GIS users wish to be able to access spatial data, and indeed the geoprocessing capabilities, located at a remote GIS that theoretically could be situated anywhere in the world.

One of the key concepts that need to be mastered in order for truly open systems to be developed is the concept of metadata. Metadata is commonly defined as "data about data" (ANZLIC 1996, ANZLIC 1997, Kildow 1996, Shelley and Johnson 1995). There are two different forms of metadata. The first, and oldest, form of metadata occurs within GIS, CAD packages and databases where it is "[an] underlying set of rules which tells a software program how to handle data" (Wilson 1998). Database management systems have for a long time used metadata to describe the internal layout of the data schemes within them (Codd 1990, Korth and Silberschatz 1991). This description of the internal schemes is then used by the database management system to construct the results of user queries. By using the metadata the system knows where to find the results for the query.

The second form of metadata is a recent development. Metadata have become products in their own right, especially in the spatial data management field, where they are used to describe the characteristics of datasets. Characteristics like the custodian, description of the data, geographic extents of the data, currency of the data, storage format, data quality, contact information to inquire about the dataset are all described. In this context metadata is extremely important for spatial data as it allows a potential user of a dataset to determine whether the dataset is useful to them or not. Metadata systems can be established that allow users to search the metadata records for the datasets located on a network. From the results they are able to determine if there are any datasets that may be of interest to them, how to gain access to them, any constraints on using them, etc. Such an application is often referred to as a spatial data directory or in some cases a clearinghouse.

Metadata of this type are extremely important as they facilitate the more efficient use of spatial data. This is achieved by allowing potential users of spatial data to search for datasets that may suit their needs. They can look at the metadata record for a dataset and see if it meets the criteria for use that they have set. The record will also tell the searcher the access rights/constraints of the dataset. All this is very important as it is usually cheaper to purchase that spatial dataset from another party that has produced it for another purpose than it is to reproduce the dataset oneself. The last thing organisations want to do is to duplicate work that has already been completed by another party.

It is the contention of this thesis that metadata will be used in the future to describe the internal layout of a huge “networked” GIS. Within the networked GIS it could be said that each spatial dataset that exists on the network is analogous to a scheme or table in a closed GIS. The metadata that is used to describe each dataset could be used in the same way that it is used in a traditional database, or GIS, and hence be used to display spatial data from two or more datasets that are located on one or more servers within the network. Metadata will still be used as an end product in its own right, as there will still be the need to determine whether a certain dataset is relevant the first time it is encountered, however its main use will be to find and query spatial datasets across a network. This contention contrasts from the current use of metadata in the spatial data industry where it is used solely as an end product.

1.2 What is a Metadata Engine?

A metadata engine is an application that is used by database management systems (DBMS) to extract and display the results to a user's query. They work by parsing the users query and then consulting the databases data dictionary which contains metadata that outlines the databases internal structure. By comparing the parsed query with the metadata in the data directory the user's query is able to be resolved (Korth and Silberschatz 1991). The metadata engine works completely in the background with no direct interaction with the user of the database. The user of the database does not even have to know that the engine exists. All the user is concerned with is writing the query and then getting the right results returned. Obviously the type of metadata that are being referred to here are the first type mentioned previously.

Metadata engines should not be confused with metadata systems. A metadata system is very similar to a search engine. They allow a user to search metadata records, which have been produced to describe the characteristics of a dataset, and determine whether they wish to gain access to the dataset. Data directories and clearinghouses both use metadata systems to allow users to search them. They both contain databases that hold the individual metadata records for each dataset that is available. These databases are searched by keyword, geographical location, date, etc. and return the individual metadata records that satisfy the users query for them to view. By viewing these metadata records the user is able to determine whether the dataset is of use to them, whether it meets their accuracy requirements, if any access constraints exist, whom to contact to gain access to the dataset, etc. In the case of a clearinghouse there is also the capability to download the dataset online. However there is no capability to query the dataset online, whereas a metadata engine has this capability. An example of a metadata system is that of GI Connections in Victoria, located at http://www.giconnections.vic.gov.au.

At the present time there appears to be no true metadata engines in existence that allow the distributed processing of spatial data over the WWW. Distributed processing is the term used when an application is set up that allows the querying of autonomous databases that are located over a network. To the user of a system that allowed distributed processing of spatial data it would appear as if they were just using one integrated database. It should be transparent to the user that the data that returns after they submit a query is actually returning from possibly two or more autonomous databases.

1.3 Research Objectives

The primary objective of the research in this thesis is to prove the concept that distributed processing, and in particular, the display of spatial data from two or more data sources on web browser is possible. This is to be achieved via the development of a prototype that will use the public domain software package Isite as a base (Gamiel and Warnock 1994). The prototype will display spatial data, of possibly different formats, from two or three different servers on a simple spatial data viewer, developed at The University of Melbourne as a result of a concurrent research project, that can be loaded from a web browser such as Netscape or Microsoft Explorer.

By proving the concept of distributed processing via a prototype the premise that metadata, with regards to spatial data, is going to include the capability of being used as a mechanism to query and display autonomous datasets located across a network. Standalone GISs have suited the needs of most spatial data users up until this point, however it has now got to a stage where distributed processing of spatial data has become essential for the development of the spatial data industry as a whole.

In the past when a user of a GIS wanted to query two distributed spatial data sets the user would have to firstly integrate the two spatial data sets into the one GIS. The incompatibilities of many data sets, due to their file formats being different, meant that this was sometimes an extremely difficult and time consuming task. With distributed processing the need to merge two data sets into the one GIS before querying is eliminated. The two distinct spatial data sets could instead be queried across a network and the results merged using a metadata engine.

1.4 Research Process

The research undertaken in this project is by no means being conducted in isolation. The research is being conducted as part of a larger research group at The University of Melbourne. The group has as one of it's main objectives to expand it's understanding of SDI development. A member of the group is already at an advanced stage of examining SDI development at a local government level. A new member of the group is planning a PhD project examining the development of multinational/regional SDIs. The group is also currently seeking a 3 year funding from the Australian Research Council to study the dynamics of development of the national SDI using the DCDB as an example.

Apart from its expertise in SDI development, the research group is examining the problem of upgrading and updating SDIs using the experience of jurisdictional DCDB. One of its members, Iestyn Polley, has developed a prototype application to demonstrate the feasibility of transferring data to and from datasets, using the Internet, within these SDIs. As part of this he has developed a web application that is capable of viewing and updating spatial datasets across a network. This spatial data viewing capability is to be integrated with the concepts in this thesis to create an application that is capable of viewing many different datasets located on many different servers at the same time.

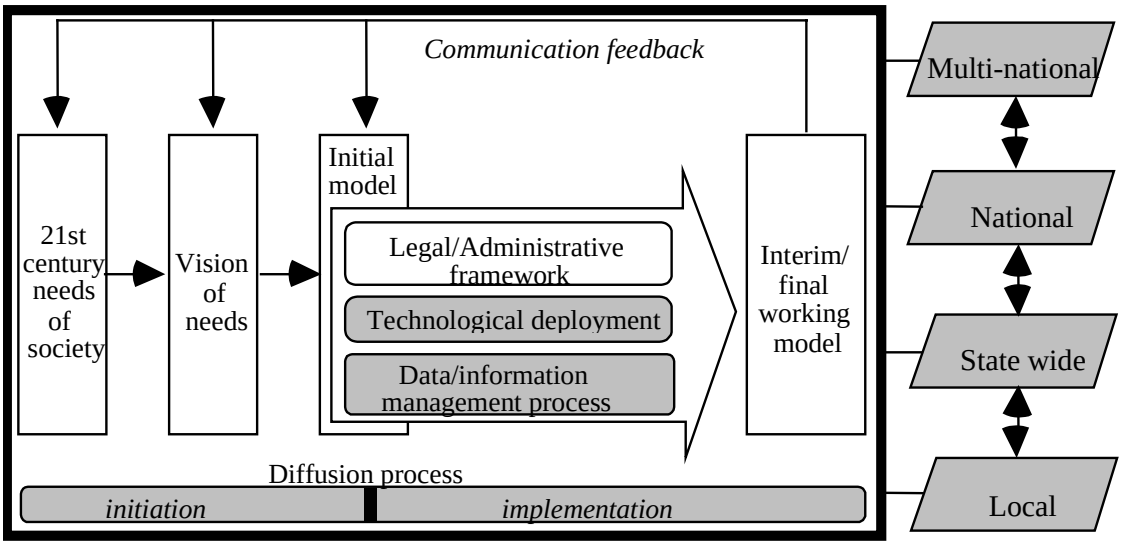

Figure 1-1: Multi-level integrated research.

Figure 1-1 depicts the multi-leveled integrated research framework that the group is participating in. This project investigates the technical deployment and data/information management processes of distributed dataset access across the Internet.

The research process that is to be undertaken in this thesis adheres to the following five steps:

- Review the developments in SDIs, OpenGIS, distributed information retrieval, as well as other projects that are underway in the area of interest of this thesis.

- Develop data models for a metadata engine that will allow distributed processing. This stage involves the development of various different alternatives for data models that will be able to facilitate distributed processing. From the options that are developed in this stage a prototype will be developed in order to provide proof of concept.

- Develop a prototype that gives proof of concept for distributed processing using one of the models developed earlier. It should be pointed out here that the prototype that will be developed is not aimed at being to be a “marketable” product, but rather a proof of concept for the theories developed in this thesis.

- Test and gain feedback on the prototype. This will involve the distribution of the prototype to various people who have been involved in supplying information and knowledge to the thesis. Feedback will be obtained on the prototype and its limitations. It is from the feedback obtained from this stage that the conclusions for the thesis will be obtained.

- Make conclusions about the prototype metadata engine and the concept as a whole, and write the thesis. The last stage in all thesis is to actually write it and make conclusions about the work that has been done. This thesis is no exception.

1.5 Justification

There are many reasons why research in the area of metadata, metadata engines and the WWW is justified. This section of the thesis will attempt to outline just a few of them with regards to the Department of Infrastructures (the sponsor of this thesis) needs and the community as a whole.

The first, and most important, reason for conducting research in this area is the fact that duplicated data collection is a waste of time, effort, and money. One of the areas in the spatial data industry where a lot of money can be saved is in reducing the amount of spatial data that is collected two or more times. This can be achieved by making it known to all users, or potential users, what spatial data are already in existence. One of the ways to do this is through metadata. Metadata which can be used to construct spatial data directories that give details of the spatial data that is available on the network. Metadata helps in the eradication of duplicated effort through the identification of spatial data sets that are of use. A spatial data set may as well not exist if people don't know that it does.

The second reason for research in this area is that the sharing and/or selling of spatial data can be a means of recuperating some of the expenses of obtaining the dataset. Using metadata allows other users of spatial data to know what spatial data is available. This means that there is the real possibility that some of the money spent in acquiring the dataset can be recouped via data sharing/selling with other organisations that may have otherwise had to collect the data themselves. Both the seller and purchaser should benefit in this situation, as the data should be cheaper to buy that collect oneself.

A third reason and the most important for research into the areas of metadata, metadata engines and SDIs is that combining spatial datasets can be a great way to obtain new "information". Information is obtained from analysing datasets and finding patterns that actually mean something. An example is when the Department of Infrastructure (DOI) studies the land development database and is able to conclude where the growth corridors in Victoria are.

Distributed processing allows the querying of two or more datasets at the same time. By using a metadata engine and live links to each of the available datasets it should be possible to interactively query one to many datasets concurrently. This may allow for datasets that have not previously been combined to be integrated and thus allow new knowledge to be obtained from these datasets.

With reference to the Department of Infrastructure it has many different datasets, in many different formats, spread across the entire organisation. Being able to view, query, and search all these datasets at the same time would be a great advantage to them. At the current time most of the agencies within the department work independently and look after their own data. If the DOI had the ability to perform distributed processing on its datasets then it is highly likely that they would be able to perform the task more efficiently, and in fact may be able to provide more services to the community through the acquisition of new knowledge.

1.6 Structure of the Thesis

This thesis has seven chapters, including the introduction (chapter 1) and conclusion (chapter 7), and will be set out as follows:

Chapter 2: Relevant Concepts. This chapter will define in some detail the general concepts of:

- Spatial Data Infrastructure;
- Data Warehouses;
- Data Marts;
- Data Clearinghouses;
- Data Mining; and
- Interoperability, in terms of
    - SDTS and OGIS.

These concepts are all relevant to the general research area undertaken in this thesis and it is worthwhile to define each of these concepts to gain an in depth knowledge of them.

Chapter 3: The Current Situation. Whereas chapter 2 outlines some of the general concepts involved in spatial data and information handling, this chapter will describe in detail some of the most relevant developments that have been occurring throughout the world. The chapter is divided into three sections, being Victorian, National, and International developments. Developments that will be outlined include the:

- Victoria's Geospatial Information Strategy;
- Land Channel;
- Department of Infrastructure GIS Strategy;
- GI Connections;
- Basic Land Information Network (BLIN);
- ANZLIC Metadata Guidelines;
- Australian Spatial Data Directory; and the
- New Brunswick Real Property Information Internet Service.

Chapter 4: The Metadata Engine - Concepts. The previous two chapters went into some detail about related concepts and projects. This chapter gives a description of the concepts behind developing a metadata engine. The topics that are discussed are: the

importance of metadata engines, the users view of the data, the data models that could be used, the live connection to the data server/s, and the concepts behind displaying data from several data servers on the screen via a web server.

Chapter 5: The Metadata Engine - Developed. Whereas chapter 4 discusses the many options available for the development of a metadata engine, this chapter gives a description of the actual options used to develop the prototype metadata engine. A few of the topics that are discussed are the data model used, Isite (an Internet metadata publishing tool), the methods that were used to deal with the querying of several different data sources at once, and the modifications that were made to the Isite source code in developing the prototype.

Chapter 6: Limitations and Improvements. The prototype that is discussed in the chapter 5 has many limitations. This chapter outlines some of these limitations along with the improvements that could be made in order to overcome these limitations.

Chapter 7: Conclusions. This chapter discusses whether the objectives of the project were met along with a general summary of the whole thesis. It also discusses areas where further research is needed.