1 Relevant Concepts

1.1 Introduction

In the last decade or so the amount of spatial data that has been collected in digital form has increased dramatically due to the rapid development of spatial data capture technologies. The advancement in 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. As such, there is now a vast amount of spatial data in digital form, stored by many organisations at various locations across the globe, much of which are not being used as effectively as they should (Phillips, *et al.* 1998b).

In recent times there has been a greater focus on how best to use spatial data that has been collected and stored in large "electronic silos" (Gore 1998). The integration, and subsequent querying of spatial datasets, the locating and obtaining of datasets across a network, and the transfer of dissimilar spatial datasets across networks are all concepts that have arisen in an attempt to better utilize the spatial datasets that are in existence.

The purpose of this chapter is to describe in detail the concepts behind the transfer of spatial data across a network detailing the concepts of SDIs and interoperability – spatial data transfer standard (SDTS) and the Open Geodata Interoperability Specification (OGIS). The chapter also aims to describe the concept of the integration of datasets, data warehousing and data marts, as well as the subsequent querying of the integrated datasets – data mining. Finally the concept of locating and obtaining a dataset across a network, clearinghouses, is discussed.

1.2 Spatial Data Infrastructures

Geographic Information Systems (GIS) can be viewed as a tool, a resource, as well as part of an overall Spatial Data Infrastructure (Phillips, *et al.* 1998a). GIS as a tool is a software package that is capable of integrating spatial and non-spatial data to yield the

spatial information that is used in decision making. The database that has been created, maintained and exploited using a GIS – the tool – is itself often referred to as a GIS. Often the database consists of data that are collected for a particular project data which, are in most cases, useful for other projects. In this context GIS is a resource.

Cooperation and collaboration of several disciplines and a proper strategic plan are usually required in the maintenance of these GIS resources. The resources are maintained at the state or national level, and sometimes by private corporations. Coordinating authorities are needed in such cases with different users (which may be agencies) being assigned custodianship and usage privileges for subsets of the data. Users in the general community are then able to expect the data to be available, and with network technology, to be transparently accessible. At this level, the GIS have now acquired the status of an infrastructure–the spatial data infrastructure (SDI) (Phillips, *et al.* 1998a).

SDIs are just like other forms of better-known infrastructure, such as roads, powerlines and railway. The whole concept of SDIs, and other forms of infrastructure, is that they allow authorised and/or participating members of the community to use them. They are simply available and taken for granted, although we may pay for the right to use them, for example through vehicle registration, railway tickets etc. Users essentially do not care how they work or who makes them work (Phillips, *et al.* 1998a).

SDIs comprise the fundamental datasets (spatial data resource) as well as the interrelationships between these datasets, the management of them, and the means of access to, and distribution of, those data. The FGDC (1996a) defined an SDI as an "umbrella of policies, standards, and procedures under which organisations and technologies interact to foster more efficient use, management, and production of geospatial data." It further explained that it "consists of organisations and individuals who generate or use geospatial data, of the technologies that facilitate use and transfer of geospatial data, and of the actual data." It should at no stage be assumed that SDIs are all about networks and technology (FGDC 1996a, Masser 1998). An SDI will not

function, no matter how good the networking and technology is if communication channels, standards, procedures, partnerships and data have not been developed.

SDIs allow the sharing of data. This is extremely useful, as it enables spatial data users to save money, time and effort when trying to acquire new datasets. This is important, not only to the organisations looking for the data, but also for the custodians of the data. Due to the "commoditisation" of data, custodians can use the SDIs to attempt to recoup some of the production costs of the dataset by selling/trading/sharing it with other organisations. They also help to minimise the duplication and fragmentation of fundamental datasets that have already been captured at great expense (Mooney and Grant 1997).

1.2.1 Data

The actual spatial data that reside in an SDI are obviously the most important component in an SDI. SDIs cannot exist without spatial data. For a spatial data resource to acquire an infrastructure status it needs to develop to a stage where it is a dataset that is accurate, up to date, consistent, updated in one place only to avoid duplicate datasets arising, and used by members of the spatial data community as, essentially, a base dataset that other spatial data overlaid upon. Prime examples of datasets that have, or are acquiring, an infrastructure status are the cadastral and topographic databases although they require a great deal of effort in getting them up to date, consistent and accurate (Phillips, *et al.* 1998a).

When an SDI is being developed, one of the biggest issues that has to be addressed is what to do with all the legacy data (data that are already in existence) (McKee 1996). The larger the SDI that is being developed, the greater the amount of legacy data that are going to be present. The problem with legacy data is that they are likely to be stored in all sorts of different proprietary formats, making the sharing/selling of that data difficult if the purchaser of the data uses a different proprietary data format. Legacy data tends to be mainly project-specific data, since for data to acquire an infrastructure status it should be in a form that is easily transferred form proprietary data format to another (Phillips, *et al.* 1998a).

1.2.2 Communication

The second aspect of an SDI is communication. At its most fundamental level an SDI consists of the individuals who are concerned with spatial data, both users and producers. One of the most important first steps in the creation of an effective SDI is the establishment of good communication channels between people/organisations concerned with spatial data (FGDC 1996a). The development of good communication channels between individuals and agencies within the spatial data community allows for the establishment of partnerships, standards and procedures. These in turn allow for data to be shared/traded/purchased amongst the different data custodians (Phillips, *et al.* 1998a).

1.2.3 Common Standards and Procedures

The third aspect of an SDI, which is brought about with the aid of good communication channels, is the introduction of common procedures and standards. Common procedures and standards facilitate the sharing of data across the spatial data community to a greater extent. An analogy can be drawn with the transportation infrastructure. In the transport infrastructure, standards dealing with rail gauges, road sizes, and the side of the road to drive on, are just a few of the many standards that are in place to help people make better use of the infrastructure. This is similar with Spatial Data Infrastructures. Having datasets in an SDI that are stored in different formats means that the sharing of these datasets is difficult due to the many incompatibilities that exist between the datasets. Many software products will not read data made by other software products, and hence the best utilisation of the data cannot be obtained. By having standards for data storage etc. in SDIs, data can easily be shared amongst users and the best possible utilisation of the data can be achieved (Phillips, *et al.* 1998a).

Common standards within an SDI tend to solve many of the incompatibility problems for newly created data, however the legacy data will remain a problem. Many organisations have significant amounts of money tied up in systems that have legacy data that are not compatible with other legacy data used by other organisations.Very few of these organisations are willing to sacrifice their own investment in order to

have an effective SDI (Phillips, *et al.* 1998a). Two standards which benefit SDIs are the Spatial Data Transfer Standard and the Open Geodata Interoperability Specification. These standards will be explained later in this thesis.

1.2.4 Partnerships

A fourth aspect in the development of SDIs is in the establishment of partnerships for the transfer of spatial data and establishment of common databases. Partnerships are a major achievement in the establishment of an SDI since it is often seen by companies as giving up their competitive edge to share, trade, sell and create data with other companies (FGDC 1996a). Partnerships are extremely important. A good network of metadata and transfer standards enabling users to see what data are available is useless if the custodians of the data are not willing to share or sell their data.

1.2.5 Technology

The fifth, and final, aspect of an SDI is the actual technology that is involved. There are two aspects to the technology aspect of SDIs. The first aspect is the actual technology that deals with communicating over networks. Much of the SDI technical infrastructure dealing with networks has already been built, or is in the process of being built, by the world's information technology (IT) industry as they build a global information infrastructure. Computers keep getting faster and the telecommunications and distributed computing hardware/software/standards infrastructure which supports distributed geoprocessing is spreading at a rapid rate (McKee 1996).

The second aspect to the technology component of SDIs is the technology that is required to allow data to acquire an infrastructure status. It is not only the spatial data capture technologies that are important, but also the data models that have to be developed in order to make the dataset as portable as possible. In Victoria work is already underway to allow many key datasets (cadastral, topographic etc) to acquire an infrastructure status (Phillips, *et al.* 1998a).

1.3 Interoperability

Interoperability is defined by the Open GIS Consortium (OGC) as being the ability to:

> 1) freely exchange all kinds of spatial information about the Earth and about objects and phenomena on, above, and below the Earth's surface; and 2) cooperatively, over networks, run software capable of manipulating such information. (Buehler and McKee 1996)

In other words interoperability is the ability to be able to exchange and manipulate spatial data across wide area networks without having to consider the format of the data or the system that is manipulating them. An example of what is possible when systems are interoperable occurs when a user, who is using a web browser, is able to do the following:

1) Transfer spatial data from a remote GIS to their own local GIS and start using the data immediately, even if the data in the remote GIS and the local GIS are not of the same format.

2) Manipulate data that are stored in the user's local GIS with tools that are located in a GIS that is at a remote site. In other words the user can download functionality from the remote GIS into their own GIS.

3) Manipulate data that are stored at a remote GIS with tools that are located at the user's local GIS. In all cases it should not matter what format the data are stored in the remote or local GIS (Phillips, *et al.* 1998a).

Two attempts have been made at achieving interoperability. These are the Spatial Data Transfer Standard (SDTS) and the Open Geodata Interoperability Specification (OGIS). SDTS is more about data sharing rather than the 'operability' in the fullest sense.

1.3.1 The Spatial Data Transfer Standard

GIS packages use proprietary storage structures and formats. Data sharing involves downloading data from one GIS and translating them into the format of the target

system before loading that data into that target system. Vendors have developed “bi-lingual” or “bi-lateral” translators between pairs of software systems. To exchange data between n different cooperating systems would require n(n-1) such translators or $\frac{n(n-1)}{2}$, if each translator is bi-directional. However, using an intermediate transfer format, the number of translators required reduces to n, each system providing a translator between its proprietary format and the adopted standard. In CAD applications, 'dxf' has become an industry standard. However, while the dxf format can be used to transfer the geometrical component of spatial data, it does not provide for topological data and other cartographic features. Therefore there is a need for a transfer standard, similar to dxf, but with provision for cartographic features (Phillips, *et al.* 1998a).

The Spatial Data Transfer Standard (SDTS) is an intermediate transfer format for the transfer of spatial data. As SDTS is designed to support any type of spatial data, implementing all of its options at one time is impractical due to the sheer size and complexity of the transfer file that would be created (Lazar 1992). Instead the SDTS is implemented through the use of Profiles. A Profile is a well defined subset of SDTS created for specific types or models of spatial data and uses as few of the SDTS options as possible. Due to the fact that a profile is meant to transfer a specific type of spatial data, it identifies only the portion of the SDTS that applies to that data model and excludes all other parts. For example the topological vector profile (TVP) is designed as the intermediate transfer format for topologically structured vector data. In the TVP the SDTS portions that deal with raster data are not included in that profile (USGS 1996).

In order to transfer data between systems that use differing proprietary formats using profiles, a translator, or spatial data transfer processor (SDTP), has to be used to encode and decode the spatial data to and from the profile (Althiede 1992a). Encoding is the process of extracting information from the dataset(s) at the local GIS and placing it into transfer files that conform with the relevant profile. Decoding is the process of extracting information from the transferred files and placing it into dataset(s) at the target GIS (Althiede 1992b). Figure 1-1 is a simple example of a SDTP encoding spatial data from an Oracle format into an SDTS profile and then a

separate SDTP decodes the spatial data in the profile into an INGRES format. The SDTPs would normally be set up so that the encoding and decoding can occur at both ends.

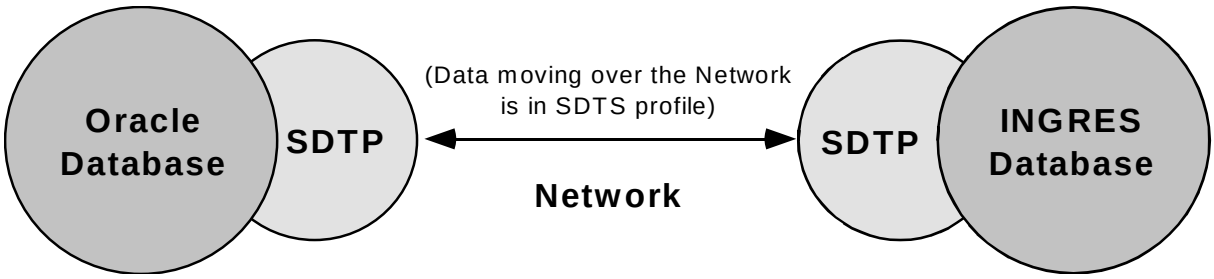

Figure 1-1: Example of a SDTS connection.

1.3.2 The Open Geodata Interoperability Specification

The Open GIS Consortium Inc. (OGC) was founded in 1994 in response to widespread recognition of incompatibilities in spatial data transfers and its many negative consequences for industry, government and academia. The members of the OGC share the positive vision of a global information infrastructure in which geodata and geoprocessing resources move freely, are fully integrated with the latest distributed computing technologies, accessible to everyone, “geoenabling” a wide variety of activities that are currently outside the domain of geoprocessing, opening new markets and giving rise to new kinds of businesses and new benefits to the public (Buehler and McKee 1996). Geoprocessing software vendors, database software vendors, visualisation software vendors, system integrators, computer vendors, telecommunication companies, universities, information providers and US federal agencies have joined the consortium to participate in creating a software specification and new business strategies that will help solve these problems and fulfil these potentials.

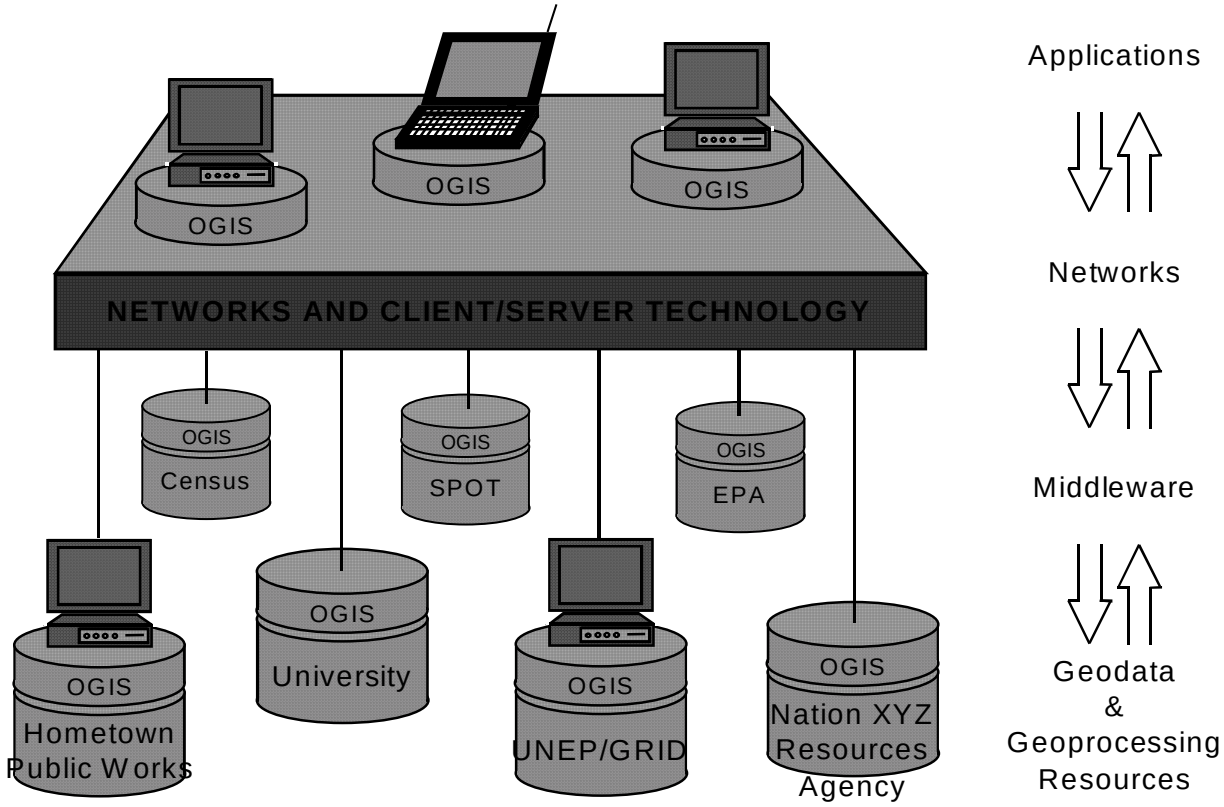

Figure 1-2: Example of what a network built using OGIS compliant software will look like. (Buehler and McKee, 1996)

As a result of the work undertaken by the consortium the Open Geodata Interoperability Specification is being developed to act as an interface between different software packages. Developers that build systems with OGIS interfaces will create middleware, componentware and applications that will be able to handle a full range of geodata types and geoprocessing functions. Users of these systems will be able to share networked data in which all geodata conforms to a generic data model, even though the data may have been produced at different times by unrelated groups using different production systems. The data may have been produced for different purposes and may in fact still reside under the primary control of the system used in its production. This geodata that remains stored in the format of the original production system, and hence is not OGIS compliant, will be accessed via software that will encapsulate the data to bring it into compliance, i.e. transform the data before transfer into OGIS, as shown in Figure 2-2.

1.3.3 Difference between SDTS and OGIS

The main difference that exists between SDTS and OGIS is that SDTS is a transfer standard, whereas OGIS is an operational standard. SDTS has only one goal, which is to provide a transfer standard for spatial data. It allows spatial data that has been captured and stored in different formats to be converted into a neutral format for transfer between differing software systems, whether they be simply on the same computer, or located over a network on two separate computers. OGIS on the other hand not only has the SDTS goal of easy data transfer, but also the goal of allowing a user on a local machine to use the geoprocessing capabilities of the remote server on the remote servers data, whilst viewing the results on their local computer. Therefore OGIS is not just a data transfer standard, but a much more complete operational standard to allow truly interoperable systems.

1.4 Data Warehousing

Data warehousing is a concept that has been around for many years for non spatial data and has recently begun to enter into the spatial data area. Brobst (1996) describes a data warehouse as:

> [a] complete repository of corporate data extracted from transaction systems that is available for ad-hoc access by knowledge workers.

Inmon (1995b) describes it as:

> the central architecture for information systems of the 1990's. [They are a] subject-oriented, integrated, time-variant, non-volatile, collection of data in support of management's decision making process.

Combining these two descriptions a working definition for a data warehouse for this thesis is

> A data warehouse is a subject-oriented, integrated, time-variant, non-volatile, collection of data that has been extracted from transaction systems and is used by knowledge workers to support decision making.

1.4.1 The Operational and Data Warehousing Environments

It is important to note at this point that in the data storage and analysis world there are two main types of environments that are used within a company. These are:

- The operational/application environment; and
- The data warehouse environment (Brobst 1996, Orr 1996).

Some of the differences that exist between these two types of systems are summarised in Table 1-1.

Table 1-1: Differences between the data warehouse environments and the operational/application environments.

Operational Environment	**Data Warehouse Environment**
An operational system runs the business.	A data warehouse gives an awareness into how to improve the business.
An operational system is designed around the applications and functions that a business has to perform.	A data warehouse is designed around the major subjects that a business is involved in. It is subject-oriented.
An operational system is concerned with both database design and process design.	A data warehouse is concerned with data modeling and database design exclusively.
Operational data contain data to satisfy immediate functional /processing requirements.	Data warehouse data exclude data that will not be used for decision support system (DSS) processing.
Operational data maintain an ongoing relationship with two or more tables based on a business rule.	Data warehouse data span a spectrum of time and the relationships within the warehouse are many.

To better emphasis these differences an example of where these systems could be used within an organisation is given. The Intergraph ambulance dispatch system that operates in Victoria is a good example of an operational system. In this case the operational system has data in it that represents the current situation in terms of road networks, street addresses and ambulance locations. These data are used by the system to dispatch the closest ambulance(s) to an emergency. Essentially the system runs the whole ambulance dispatch service. Such a system would typically not have the ability to analyse any growth areas in ambulance requirements (Phillips, *et al.* 1998a).

A data warehouse as listed in Table 1 is defined as giving an organisation an awareness into how to improve its business. The same organisation that uses the dispatch system to dispatch ambulances to emergencies could also use a data warehouse to analyse trends in ambulance dispatches throughout the catchment area. The data that are in the data warehouse do not reflect the road networks, street addresses and ambulance locations at the current time, but where ambulances have been dispatched to and from over a period of years. If data such as population growth, age profile and planning data were added to this dispatch data it could be used to predict where new ambulance stations could be placed and the number of ambulances allocated to each station to provide an optimum service. The data warehouse has the ability to analyse these data and helps in finding the trends that are contained within it (Phillips, *et al.* 1998a).

1.4.2 Properties of a Data Warehouse

One of the most important properties to note about a data warehouse is that data entering the data warehouse, in almost every case, are extracted from the operational systems. The data are taken/copied from the operational system, transformed and stored in a physically separate store. Integration is a key part of the transformation process. Without exception all the data that are located within a data warehouse are integrated. This is the most important aspect of the data warehouse environment (Inmon 1995b).

Another property of a data warehouse is that it is time variant. All the data that are stored within the data warehouse are accurate as of some moment in time (Inmon

1995b). The time variance nature of the data means that all data entries in the warehouse must have a time stamp associated with them. This stamp forms part of the key for that entry. The data that are stored in the data warehouse do not have to reflect the current situation, but instead a situation that has occurred in the past. This is very different from the data that are found within an operational system. Operational system data are accurate for the time that the system is accessed. The operational environment reflects the current situation, whereas data warehouse data represent a situation in the past.

Another point that should be made about data warehouses is that they are non volatile. Once the data have been entered into the data warehouse correctly they do not get removed. Only three kinds of operations occur in a data warehouse, the initial data loading, user querying of the data and summation of the data. The structure of the data warehouse is such that there are different levels of data detail. Data that have recently been loaded into the warehouse are very detailed, whereas data that are, for example, 10 years old are often highly summarised, due to their lessening relevance, with varying degrees of summarisation between. Summation is the process of reducing detail. For example data that are less than a year old may have an entry for each day, whereas data that are 10 years old may have one single entry summarising the whole year. The rest of the data, for that year, are either discarded or archived, depending on the organisation.

1.5 Data Marts

A data mart is a departmental data warehouse. The data that are located in the data mart have been extracted from the corporation’s data warehouse and customised for the departments own decision support system (DSS) usage. Thus “a data mart is a body of DSS data for a department that has an architectural foundation of a data warehouse” (Inmon 1995a).

The type of data that reside in a data mart are slightly different from the data that reside in a data warehouse. The data that reside in a data warehouse are at a granular level, whereas the data that reside in the data mart are more refined. The different data

marts that exist within a corporation contain different combinations and selections of the same detailed data that are in existence in the corporations data warehouse.

It may be asked at this stage why data marts are necessary within a corporation that already has a data warehouse? The main reason for this is the consequences of the corporations data warehouse growing very large very quickly. Data marts, due to the fact that they contain only selections of the data that exist within the data warehouse do not have this problem. Some of the consequences of a data warehouse becoming excessively large include (Inmon 1995a):

- The competition between the DSS analysts within the corporation becomes fierce. Increasingly departmental DSS processing is done within the warehouse to the stage where the warehouse can no longer handle the load.

- As the volume of data within the data warehouse increases, so too does the cost of doing the processing.

- The elegance of the software that is available for handling the data within a data warehouse decreases dramatically as the volume of data within increases.

- The more data that there are, the harder it is to customise the data into the form that is required. When there is a small amount of data in the data warehouse, the DSS analyst is able to customise and summarise the data every time an analysis is undertaken. When the volume of data is large, this process simply takes too long.

Once the corporation's data warehouse reaches a certain size it is much more attractive to start to develop data marts which have much less data stored in them. This approach offers the following advantages:

- Each department can customise the data from the data warehouse as it flows into its own data mart. As there is no need for the data in the data mart to serve the whole of the corporation, the department can summarise, sort, select, structure, etc., the data to the form that it believes bests suits it without having to consider the needs of other departments.

- The amount of historical data that are needed in the data mart is a function of the department and not the corporation. This often means that the department can select a much smaller time scale of data than that of which is found in the data warehouse.

- The department can do any type of DSS processing at any time without having to consider what effect this processing may have on the other departments in the corporation. As all the data marts are separate entities what one department is doing has no effect on the efficiency etc., of the other departments.

- Each of the departments can select the software for its data marts that best suits its needs. Each of the departments within a corporation may undertake DSS analysis using data marts that run using different.

- DSS analysis costs less, per unit, to undertake on smaller datasets than it does on larger datasets. This results in significant cost savings over the long term (Inmon 1995a).

An example of where a data mart would be of particular use would be in a government organisation such as the Department of Infrastructure (DOI) in Victoria. The DOI was established as a result of bringing together business units and agencies with an "Infrastructure" focus. Each of these business units and agencies bought their own data into the merger and use it for different purposes. It is in this situation that data marts can be used to their full potential. All the data could be stored in one large data warehouse which could be analysed by people wishing to make decisions with a bias towards the whole of the organisation. The merging of all the datasets from all the different agencies and business units could make it possible to see trends in the phenomena the data represents that may not be seen from the individual datasets. Having all the data in one data warehouse is reasonable when there are few people wishing to use the warehouse or wanting to undertake analyses with a single business unit or agency bias. The problem with the DOI example is that there are many people wishing to undertake analyses with a business unit or agency bias and hence they slow down the system for everyone. If each of the business units had their own data mart

set up the way that they required to make best use of their data, performance of the department could be increased significantly.

1.6 Data Mining

The world has seen a dramatic increase in the amount of information or data that are being stored in digital format over the past 20 years. This accumulation of data has taken place at an exponential rate with the amount of data in the world doubling ever 20 months (Dilly 1995). Storing this data has become easier as large amounts of computing power became available at low cost.

After concentrating so much attention on collecting and storing of data for many years, the next problem facing many organisations was what to do with the resource. It has been recognised for many years that information is the heart of business operations and decision makers could make use of the data stored to gain an insight into the business (Dilly 1995). Analysing data can provide further knowledge about a business by going beyond the data explicitly stored to derive "knowledge" about the business. In this case knowledge is data that have been structured into a format that is actually meaningful to the observer. Data Mining is the term that is used when describing the process of deriving knowledge from data. (Frawley, *et al*. 1991) states that:

> data mining, or knowledge discovery in databases (KDD) as it is also known, is the non-trivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarisation, learning classification rules, finding dependency networks, analysing changes, and detecting anomalies.

(Holsheimer and Siebes 1994) described data mining as:

> the search for relationships and global patterns that exist in large databases but are hidden among the vast amounts of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the

> objects in the database and, if the database is a faithful mirror, of the real world registered by the database.

Summarising the above quotes to gain a working definition for data mining, it is possible to define data mining as

> the analysis of data using software techniques to find patterns and regularities in sets of data. The patterns and regularities that are found represent valuable knowledge about the data sets and the objects in them.

The underlying features and rules that are in the data are found via the software on a person's computer. By data mining it is possible to “strike gold” in unexpected places as the data mining software extracts patterns not previously visible, or in fact they are so obvious that no one has noticed them before (Dilly 1995).

There are five stages involved in the process of data mining (Dilly 1995):

- Selection – This is the process of selecting/segmenting the data according to some criteria. This way the data are broken up into subsets. For example selecting two subsets based on whether a person owns a house or not.

- Preprocessing – During this stage the data are cleaned. This involves the removal of data which are seen to be unnecessary and may slow down the queries. An example of this is where a sex attribute can be removed when we are studying testicular cancer. Integration is also conducted during this stage so that all the data are in a consistent format.

- Transformation – In this stage overlays are added to the data, such as demographic overlays, that are commonly used in market research. The data are then made useable and navigable.

- Data mining – This is the stage where the extraction of patterns from data occurs.

- Interpretation and evaluation – This stage involves the patterns that were identified by the system being interpreted into knowledge which can be used to support the human decision making process.

The users of the data mart and data warehouse environments who undertake data mining are called the DSS analysts (Inmon 1995a). These users are individuals who make strategic decisions with a departmental bias in the case of data marts and a company bias in the case of data warehouses. They are business people not technicians. DSS analysts can be divided into two categories (Inmon 1995a) :

- "Farmers"– These DSS analysts know what they want out of the system and regularly and predictably go to the same place in the system to find it.
- "Explorers" – These DSS analysts look at data in the system in a random, somewhat sporadic, fashion. These users often find nothing as a result of their queries, however occasionally they achieve sensational results.

Farmers know that a relationship exists within a dataset and they exploit that relationship to gain answers from the system. Farmers know exactly where to go to find the answer to a query. Explorers on the other hand tend not to know what they are looking for. They try to find new relationships that may exist within the data. Thus farmers use existing relationships, whereas explorers try to find new relationships for the farmers to use.

There are far more farmers active at the data mart level than explorers. Explorers are more likely to be found at the data warehouse level where there is more raw, untouched, data (Inmon 1995a).

1.7 Clearinghouses

A clearinghouse is an application that is located on a network and is used by people who have access to the network to obtain copies of datasets that the custodian has made available on the network. Clearinghouses contain field level descriptions of the data located on the network. Essentially a clearinghouse allows a user to search a network to find out what data are located on it, and then to actually gain access to that data, subject to the constraints placed on it by the data's custodian (Phillips, *et al.* 1998a).

A key aspect of clearinghouses are the metadata systems that are used to run them. In the context of clearinghouses metadata systems organise and search metadata records that correspond to datasets that reside on the network. They allow a user of the system to search for data of a certain theme by searching their corresponding metadata records. Within these metadata records, links to either the actual dataset or an order form for the dataset are supplied so the user can gain access to the data (Phillips, *et al.* 1998a).

The clearinghouse can either be a single server that has all the metadata for the entire network, or it could be a system of decentralised servers that are located on a network. The metadata is collected by each of the participating sites in a standard format. When the metadata is in a standard format it allows consistent querying and presentation across the multiple participating sites.

The Clearinghouse Activity (located at http://www.fgdc.gov/Clearinghouse/ Clearinghouse.html), sponsored by the FGDC (USA) is a good example of a clearinghouse that is currently in operation. In this example the clearinghouse is a decentralised system of servers that contain metadata on the spatial data that are located on them. The Clearinghouse Activity uses Web technology, and the search and retrieve protocol known as ANSI Z39.50-1995 (ISO 10163-1995) for the query, search and presentation of results to the Web client (FGDC 1996b). The Z39.50 protocol includes client and server software that establishes a connection, passes a formatted query, returns the query results and finally presents the identified documents to the client in one of several formats.

The fundamental goal of the Clearinghouse Activity is to provide access to digital spatial data through the use of metadata. Sites that are participating in the clearinghouse are encouraged to provide hypertext linkages within their metadata that will enable the users to download the digital data in one or more formats onto their own machine (Nebert 1996). There are two situations that can occur where this direct download over the Internet is not possible. The first is where the dataset is too large to be efficiently and safely transferred over a network, and the second is where the data have a commercial value. In both cases this can be solved by having hypertext links to order forms which can be filled out and the data can be shipped to the consumer.

Essentially the clearinghouse is a low cost advertisement for providers of spatial data, both commercial and non-commercial, to potential customers on the Internet.

Victoria has recently activated the GI Connections web site (located at http://www.giconnections.vic.gov.au) which has a data directory which is the first stage of a clearinghouse for spatial data for the state. It provides a listing of the government's spatial datasets for the state, and who to contact to get access to a specific dataset. It, at the time of writing, has the ability to allow a user to purchase the datasets over the web.

1.8 Chapter Summary

The purpose of this chapter was to describe in detail some of the concepts that are being used in developing more efficient uses of spatial data throughout a company and community. All of the concepts relate to the research topic in some way and as can be seen, many of the concepts that have been outlined are similar, or indeed depend upon each other to function. SDIs are essentially the technology, data, communication, standards and procedures and partnerships that have been put in place to allow the dissemination of spatial data across a network, whether that network be electronic or not. The concept of SDIs is at the heart of the research that is being undertaken in this thesis. Essentially what this thesis is investigating is the SDI that is required to allow the concept of distributed processing of spatial data to occur.

Interoperability is a subsection of SDIs and is the standards and procedures that are put in place that allows for spatial data to be transferred and accessed across an electronic network. Interoperability is extremely important to the concepts that are being investigated within this thesis. Distributed processing across networks relies upon the data that is stored at the different servers being compatible. The different proprietary file formats that exist within the spatial data industry means that it is quite likely that when distributed processing is being undertaken, the spatial datasets involved are not all going to be in the same format. This is where SDTS and OGIS are extremely useful as they should allow the successful transfer of spatial data between the various servers in a suitable format.

The concepts of data warehouses and data marts involve the integration of all of an organisation's operational databases into the one database. With data warehouses and data marts all the data is stored in one location and hence data mining (the analysis of data) upon the whole of the organisation's data is possible. Data warehouses and data marts are essentially the same concept, except that data marts have a more departmental focus as compared with the data warehouses overall view. The concepts of data warehousing and data marts are not directly related to the research that is being undertaken in this thesis, however from a user's viewpoint the three concepts are practically the same. This thesis investigates allowing a user to query many different databases, located across a network, at the same time. From the users point of view it should appear that they are accessing a single database. Thus it would appear to the user that they are accessing a data warehouse/mart.