



Notation for National Integration of Geographic Data


Steve Ramroop and Richard Pascoe

Department of Information Science, University of Otago, Dunedin, New Zealand.

Phone: +64 3 479 5608 Fax: +64 3 479 8311 Email: sramroop@infoscience.otago.ac.nz

*Presented at the 10th Colloquium of the Spatial Information Research Centre,
University of Otago, New Zealand, 16-19 November, 1998*

Abstract



In this paper, the authors' main focus is on presenting a notation used to represent data integration in a Geographic Information System (GIS) environment. The problem of GIS data integration is briefly discussed and a Conceptual Model, comprising of a **Data Center** and **Data Agencies** is presented, from a national perspective. The problem of data integration is divided into manageable and focussed components and the notation proposed by Pascoe & Penny (1995) is extended to document these components. The notation attempts to present the retrieval and processing of the data starting from the user query to that of the final data destination.


Keywords and phrases: Conceptual Model; Data Center; Data Agencies; notation.

1 Introduction

As GIS functionality increases, more and more data sets are collected and used within many government and non-government organizations. Walker D.R.F & Medyckyj-Scott (1992) indicated that the collection and entry of data is very expensive, and is major cost in establishing a working GIS. Consequently the GIS community has sought alternatives to data collection, (Walker D.R.F & Medyckyj-Scott 1992, Peel 1997, Johnson B. D. & Callahan 1991, Denness 1997).

Apart from the more costly methods of original data capture, (such as photogrammetry, remote sensing, and so on), the concept of **data sharing** is seen as a better solution. Dueker & Vrana (1995) indicated that data sharing is effective when it takes place within

and among organizations which operate in common geographic areas. This is more likely to occur if organizations share similar interest and mandates. Although organisations having common geographic areas is ideal for data sharing, such an environment is unrealistic because GIS data collections have already evolved within each organization as a consequence of using data sets not common among other organizations. Furthermore, achieving the level of cooperation needed would also be difficult.



In this paper the problem of data sharing is partially addressed by presenting a conceptual model of a national distributed database environment specific to GIS use. The model comprise of two entities which are the **Data Agencies** and a central **Data Center**. In Section 4, the two entities are defined in terms of their roles in sharing GIS data. The notation defined by Pascoe & Penny (1995) is expanded in Section 5 to represent further details of data sharing in a GIS environment at a National level.

2 GIS Integration Categories

Data sharing involves the integration of the geoprocessing capabilities of GIS software and databases that builds GIS applications. Dueker & Vrana (1995) categorized integration by decomposing the meaning of *Systems Integration* into Data Integration; Application Integration; Functionality Integration; and Organizational Integration. Although the focus of research described in this paper is on data integration, in which standards for data exchange are addressed, there is clearly a need to investigate the



other three areas of integration. For example, Marr A. J. & Mann (1998) have been focussed on the 'open process' architecture which falls into the category of Functionality Integration.

Green D. R. & Corbin (1996) argued that the evolution of GIS applications will only be possible if the investment in legacy data is protected and the data providers and software suppliers fully commit themselves to the vision of having open GISs where data and process can be shared. As a consequence of the need for this open GIS environment, most GIS vendors are in the process of creating the open environment whereby users can select the process of interest. Aronoff (1993) indicated that a large proportion of the money spent in any GIS implementation lies in the initial data capture. Therefore, apart from making the 'processes' open, much more emphasis is also needed for the 'openness of data'.

3 What is Data Integration?

GIS users are aware that there are considerable benefits to be obtained from data sharing, consequently, the area of 'Data Integration' emerged as an area of increased interest by researchers. Data integration is a broad topic covering areas ranging from corporate level database configurations such as countrywide databases (Kevany 1995) to that of procedures followed during the initial data capture stages (Flowerdew 1991) such as the collection of field survey data sets.

Dueker & Vrana (1995) commented that data integration is the ability to share access to other data sources or common databases. However, data integration is much more than access to data. Data integration also involves finding the data **acceptable** to the user. Flowerdew (1991, pp. 375) defined **Data Integration** as:

'the process of making different data sets compatible with each other, so that they can reasonably be displayed on the same map so that their relationships can sensibly be analysed.'

This is a practical definition of what is desired by GIS users, however, to integrate data a series of processes

are followed for data to be, if at all, compatible with each other. The process of integrating data sets, involves, in no particular order:

- selecting data sets appropriate to the GIS application;
- transferring data sets into a common data format (without unnecessarily losing any information);
- merging data sets initially collected at various level of details (such as graphic databases with varying map scales);
- correlating descriptive data (such as census tracks) to an appropriate graphic database (if at all available); and
- addressing the reliability of the initial captured data which includes missing, undefined and generalized data sets.

Therefore, in this paper National data integration is defined as:

The transparent process of efficiently retrieving, merging, transferring and moving acceptable data sets from various data agencies having a national perspective to satisfy user needs.

As the amount of data increases, the issue of data compatibility becomes a more important consideration. An area which also influences data compatibility is the area of **Data Quality**. Tayi & Ballou (1998) defined data quality as: 'fitness for use', which implies that the concept of data quality is relative. To determine 'needed quality' is difficult when different users have different needs (Tayi & Ballou 1998); however, users need to strike a balance between the final use of the data and the quality of data available. Data quality is also an area of continued research, however, further discussions on this topic are out of the scope of this paper. Other related references on this topic are (Goodchild 1995, Ballou & Pazer 1985, Wang & Strong 1996).

In modelling the data integration process at the National level the variations in the definition of data sets must be considered. GIS users need to view the entire data integration problem Nationally and not

specific to a particular organisation. Therefore, any data integration model should present each agency existing separately, however, they are all linked through a common communication medium.

The intention is to address the present data integration scenario rather than recommending starting fresh with accepted national standards to re-capture data sets. This is an important consideration because most users have already collected large amounts of data sets which collectively was very costly and time consuming. These data sets have been satisfying the immediate needs of the collectors who in most instances are also the users of the data sets. Therefore, the intent is for each collector/user to continue with their work while some other mechanism deals with the issues relating to integrating data sets. A Data Center can provide such a mechanism.

Users would feel more comfortable and would be less reluctant at implementing a data sharing solution which is capable of complementing their present business operations and thereby satisfying their business objectives. This can be facilitated if the proposed solution is cost effective and if possible, presents itself as an *add-on product* which is easily implemented and accepted by each Data Agency. The conceptual model shown in Figure: 1 is developed with this goal in mind.

4 National Conceptual Model

In the GIS community, the goal is to share data by way of a transformation process, such as that identified by Pascoe & Penny (1990). The transformation process must be able to transform all of the data components (such as coordinates, projections, topological relations, and so on) to the final data destination if full use is to be made of the data.

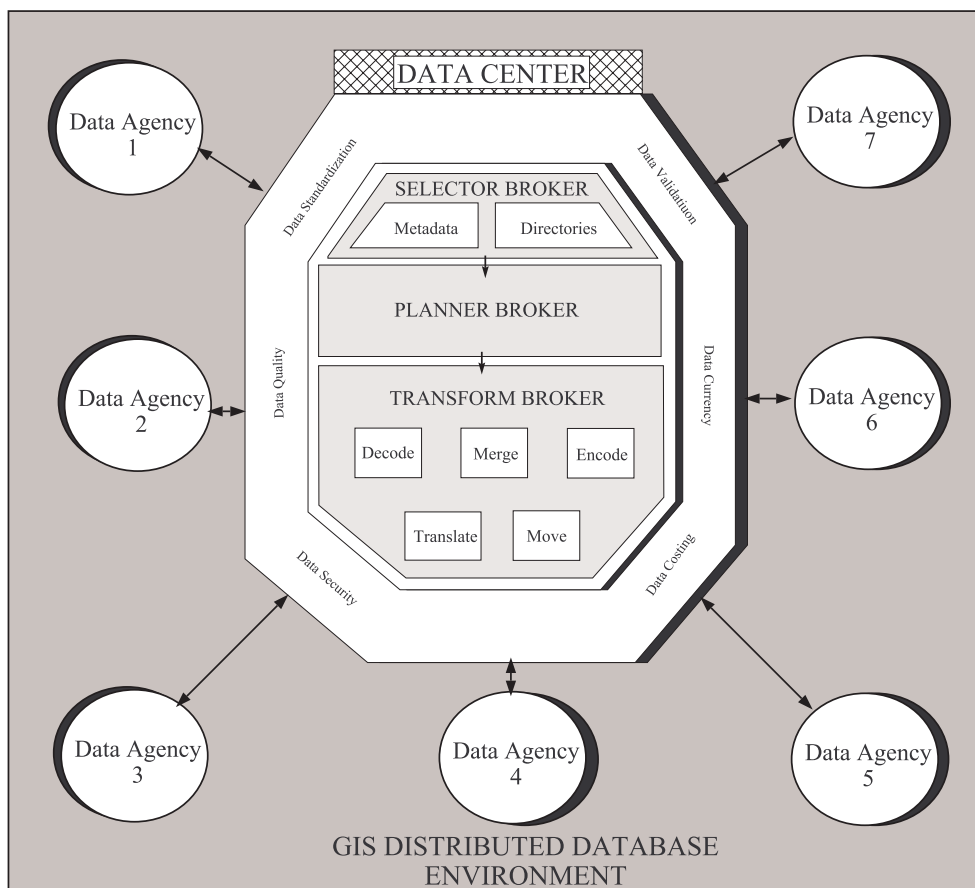


Figure: 1. National Conceptual Model



The architecture as presented in the conceptual model adopts an interfacing strategy whereby the Data Center provides a common interface. In this model each Data Agency would transfer data through the Data Center. The Data Center addresses the common processes followed in sharing data. Therefore, the notion of an interfacing strategy is redefined to include common data sharing processes which facilitate data integration at a Nation level. The model is divided into two components, which are: **Data Agencies** and **Data Center**.

4.1 Data Agencies

A Data Agency is defined as:

The producer and user of geographic data sets.

The role of Data Agencies is as follows:

- to register all data sets with the Data Center by providing meta data for each data set (for example, contact information, data format, data quality, and so on);
- to ensure data integrity by keeping data sets current and validating data sets when needed; and
- to adopt data sharing objectives common to other Data Agencies.

The model presents Data Agencies as they exist in reality that is, as separate entities whose functions vary. Each Data Agency is linked to the Data Center which communicates with the other Data Agencies. Communication is facilitated through a network connecting all Data Agencies, such as the internet.

4.2 Data Center

The Data Center is defined as:

The virtual processing center through which data sets from data agencies are located, transformed, and delivered to other data agencies who have requested data sets to satisfy their needs.

In this definition the Data Center is a **virtual** entity which is owned and governed by all Data Agencies making use of the Center's service. In essence, this creates a particular type of distributed database which is called a Federated Database System. Concepts

similar to a virtual Data Center have also been suggested by Pascoe (1996) and Abel D. J. & Tan (1998).

The definition for the Data Center depicts the Center as being the hub of activities where all issues relating to data integration are addressed. The role of the Center can be generalized as follows:

1. Assist in developing National data integration strategies for maintaining data currency, data accuracy, data costing and assessing data quality.
2. Develop databases in collaboration with other Data Agencies such that data standardization can be maintained within the context of a National Integration Strategy.
3. Encourage the use of data sets between Data Agencies.
4. Encourage further research into areas of data standardization, data quality, data validation, data security, and so on.
5. Data accessing and data processing. This includes the ability:
 - to access meta-data on the available data sets;
 - to select appropriate data sets;
 - to merge multiple data sets; and
 - to transform source data sets to multiple destination data sets.
6. Generate revenue for the Center to cover operational cost and ensure its future existence. Any cost recovery strategy would require political support, however, such discussions are out of the scope of this paper and is an area for further research.

The conceptual model in Figure 1 identifies that within the Data Center, on-going research is being done on many topics such as, Data standardization, Data Quality, Data Validation and so on. These topics are at their various stages of research (both locally and internationally), however their progress is monitored by the various Data Agencies. In effect, each Data Agency would be responsible for monitoring these areas of research as they relate to their data



sets and business functions. Once results from such researches are applied at the Data Agencies then, effectively the Data Center would be utilizing these acceptable data sets.

The Data Center is more focussed on the actual data integration processes. As such, the Center is composed of a set of 'Brokers'. A Broker is defined as:

The agent responsible for executing specific processes associated with data integration within the Data Center.

This approach is similar to that adopted by Pascoe & Penny (1995), who divided the generic process of geographical data transfers into smaller, specialised tasks which were performed by different modules. Encoder and decoder modules are for transforming the physical representation of data sets; translator modules translates data values to conform to different conceptual or implementation schemas; and communicator modules move data sets from one computer system to another via a communications network.

As will be explained in more detail in the following sections, we define:

- a Selector Broker for finding and selecting data sets appropriate to user requirements;
- a Planner Broker for optimising the data integration process overall; and
- a Transform Broker to perform the functions of the encode, decode, move (communicator), and translate modules defined by Pascoe & Penny (1995) and another new module called merge for combining data sets.

Combining the use of these brokers enables Data Agencies to retrieve and merge multiple data sets from other Data Agencies to satisfy their needs. The notation presented by Pascoe & Penny (1995) is extended in section 5 to reflect the new brokers presented in this paper.

¹ <http://www.environment.gov.au/database/metadata/anzmeta/>

² <http://www.fgdc.gov/Metadata/ContStan.html>

³ <http://www.fgdc.gov/index.html>

⁴ <http://www.usgs.gov/gils/locator.html>

⁵ <http://www.dstc.edu.au/RDU/HotMeta/qld/search.html>

⁶ <http://www.dstc.edu.au/>

4.2.1 Selector Broker

The Selector Broker makes use of metadata.

Metadata is defined as information describing the data set's location, format, accuracy, ownership, copyright, coverage, and so on. A variety of standards for such information is being defined. Examples are: the Australia New Zealand Land Information Council (ANZLIC) Metadata XML/SGML standard¹, providing guidelines for the content of geographical metadata; Content Standard for Digital Geospatial Metadata (CSDGM)², developed by The Federal Geographic Data Committee³

For easy access, metadata is stored in directories. Examples of such directories are Global Information Locator Service⁴, the Metadata search engine provided for the Queensland Government⁵ by the Distributed Systems Technology Centre⁶. Examples of tools, protocols, and standards for building directories include: X.500, an ISO-OSI standard for distributed directories; Lightweight Directory Access Protocol (LDAP), a protocol for accessing directories.

The Data Agency specifies the data to be retrieved by the Data Center in terms of a query. A query consists of values for characteristics associated with:

- data content, that is characteristics such as coverage, themes, scale, and so on;
- data access, that is maximum download time, preferred location of data, authorization, and so on; and
- representation of data, either that in which the original data sets are stored, and/or that in which the resultant data set(s) are to be returned to the Data Agency.

Using metadata and Data Agency queries, the Selector Broker is responsible for:

- extracting appropriate metadata from various geographical data directories in accordance with a Data Agency's query;
- ordering the metadata according to the preferences given with the Data Agency's query; and
- interacting with the Data Agency to select which data sets corresponding to the metadata are to be retrieved.



To illustrate the role of the Selector Broker consider the following example. A Civil Engineering Department is interested with road networks for Country 1 at a *preferred* scale of 1:2500 although larger scales will be considered. The Department formulates the corresponding query to convey this information and sends the query to the Selector Broker. Using various data directories, the Selector Broker searches for metadata describing data sets that match up with the information in the Department's query. The result of this search is a listing of all metadata satisfying the criteria, such as the following:

- (a) Road network at 1:2500 for Country 1 at the Surveying Department
- (b) Road network at 1:2000 for Country 1 at the Planning Division
- (c) Lot layout at 1:1200 for Country 1 at the Lands Registry

The Selector Broker is also responsible for prioritizing the metadata. In doing so, the Selector Broker recommends which of the possible data sets is most suitable according to information within the Data Agency's query. This recommendation is intended to aid the Data Agency in selecting the desired data set(s).

Continuing with the Civil Engineering Department example above, the Selector Broker orders the metadata listed by the Selector Broker according to the departments preferences for scale. The Selector Broker specifies this ordering as a percentage of the criteria forming the query that is satisfied by metadata for each data set. The metadata listed above are prioritized, from highest (100%) to lowest (0%), by the Selector Broker as follows:

- (a) Road network at 1:2500 for Country 1 at the Surveying Department - 100%
- (b) Road network at 1:2000 for Country 1 at the Planning Division - 75%
- (c) Lot layout at 1:1200 for Country 1 at the Lands Registry - 25%

Based upon the prioritizing of the data sets, the Data Agency selects the desired data set(s). The selection must be done by Data Agencies because they may select a lower priority data set because of information within the metadata which was not considered while they formulated their query. Once selected, the Selector Broker passes the metadata to the Planner Broker in order to construct the optimal sequence of transformations needed to produce the data matching the query.

4.2.2 Planner Broker

The Planner Broker is responsible for determining an optimal sequence of transformations needed to reorganise the selected data sets into the form required by the Data Agency. Issues considered by the Planner Broker include:

- advice from the Data Agency to assist the Planner in constructing the optimal sequence of transformations;
- source representation and location of each data set;
- representation of data set to be sent to the Data Agency;
- representation translations defined within the Transform Broker;
- time for receiving each data set into the Data Center through the communications network; and so on.

4.2.3 Transform Broker

The Transform Broker is responsible for the transformation of the selected data sets obtained from the Selector Broker to conform to different data schemas. The Transform Broker comprise of four types of modules: decoders, encoders, communicators, and translators as defined by Pascoe & Penny (1995) and merger modules to be discussed below. Each of these carry out different functions but are however related to each other in terms of their input and output.

Decoders and encoders transform the physical representation of data sets. That is, decode data values in a file format into equivalent values represented in memory and vice versa (encode). Translators map values conforming to one schema into equivalent

values conforming to some other schema. In essence this is a process of mapping values in one user context into another set of values for another user context. Communicator modules are responsible for the movement of data from one location to another through a communications network. In doing so, the communicator module resolves issues such as network types (such as internet and intranet), remote access, connectivity using protocols, and so on.

Central to data integration is the combining or integrating of related data sets. For example, zonal maps can be integrated with cadastral maps for the purpose of urban forecasting. This process is referred to here as merging and a module specifically intended for accomplishing this task, called a *merger module*, forms a part of the Transform Broker.

A transfer of a data set from a single Data Agency through the Data Center to another Data Agency will involve some combination of decode, translate, and encode modules (typically in that order). When more than one data set is to be integrated as requested by the user, the merge module will be used in combination with the other modules to achieve the transfer. The exact combination of modules and the order in which they are used is determined by Transform Broker in both cases.

4.2.4 Summary

In summary, the Selector Broker finds metadata matching a query and orders this in accordance with the preferences of the requesting Data Agency, who makes the final decision as to what data sets are to be returned. Metadata for the selected data sets are then sent to the Planner Broker to construct the optimal sequence of transformations to be performed by the Transform Broker to convert the data into a representation desired by the Data Agency.

5 Data Processing Notation

Pascoe & Penny (1995) defined a notation for describing the transfer of data from one GIS to another as a sequence of transformations. Each either contributes to the transformation of data values from a representation required by one GIS to that required by another GIS, or changes the location where data is stored. This notation is summarised in Table 1.

Extensions to this notation are now defined to reflect the additional operations (select, plan, and merge), discussed in Section 4.2, that are needed to integrate many source data sets which may be located at different Data Agencies.

5.1 Select Operator

Each data set S_i is located at a Data Agency. Using Pascoe & Penny's (1995) notation, a data set S_i

Notation	Associated term	Definition
\mapsto	A data <i>transfer</i>	The transfer of a data set from one representation to some other representation.
$\xrightarrow{\text{translate}}$	A data <i>translation</i>	The transformation of a data set to conform to a different conceptual, implementation, or physical schema.
$\xrightarrow{\text{move}}$	A data <i>movement</i>	The physical movement of a set of values from computer system x to computer system y .
$\xrightarrow{*}$	Many data transformations	One or more data translations or data movements.
Q	User query	User request specifying multiple criteria $(\alpha, \beta, \gamma \dots)$
μ	Metadata	Information associated with each data set.
$\xrightarrow{\text{select}}$	Select operator	A process which selects and prioritized metadata.
$\tau()$	Planner Function	A function representing the optimal sequence of data transformations.
$\xrightarrow{\text{plan}}$	Plan operator	An operator which selects the order and sequence of data transformations.
$\xrightarrow{\text{decode}}$	Decode operator	An operator which transforms data values in a file format into equivalent values in memory.
$\xrightarrow{\text{encode}}$	Encode operator	An operator which transforms memory data values into equivalent data values in a file format.
$\parallel \xrightarrow{\text{merge}}$	Merge operator	An operator which combines two or more data sets into one data set.

Table 1. The notation for describing the process of transferring data (Pascoe & Penny 1995)



located at Data Agency j , is denoted by $S_i(L_j)$. Thus, k data sets located at Data Agency L_j are denoted as $\{S_i(L_j):1 \leq i \leq k\}$. Not all data sets will be located at the same Data Agency. Thus, m data sets collectively located at n Data Agencies will be denoted by: $\{S_i(L_j):1 \leq i \leq m, 1 \leq j \leq n\}$. Associated with every data set S_i , there is metadata μ_i which is provided within a directory.

Data Agencies formulate a query, Q , made up of criteria $\alpha, \beta, \gamma, \dots$ specifying the nature of the data to be retrieved by the Data Center. Query Q is sent to the Selector Broker within the Data Center to produce a list of metadata for review by the Data Agency. The result of this review will be a collection of metadata $\{\mu_i:1 \leq i \leq m\}$ corresponding to the m data sets selected by the Data Agency. The notation for describing this process is: $Q_{(\alpha, \beta, \gamma, \dots)} \xrightarrow{\text{select}} \{\mu_i:1 \leq i \leq m\}$

which is abbreviated to $Q \xrightarrow{\text{select}} \mu_*$

5.2 Plan Operator

The metadata μ_* is processed by the Planner Broker as described earlier in Section 4.2.2 to produce an optimal sequence of data transformations. This sequence is treated as a function $\tau()$ which transforms the data set(s) S_* into the data set D to be sent to the Data Agency as the result of their query. The process of constructing the function $\tau()$ is denoted by:

$$\mu_* \xrightarrow{\text{plan}} \tau(S_*)$$

For q resulting destination data sets, the transformation of $\{S_i:1 \leq i \leq m\}$ data sets is represented by:

$$\{S_i:1 \leq i \leq m\} \xrightarrow{*} \{D_i:1 \leq i \leq q\}$$

where $q \leq m$ because after the transformation data sets are usually merged together. This is abbreviated to: $S_* \xrightarrow{*} D_*$

5.3 Merge Operator

Within the Transform Broker there is the merge operator, which combines two or more data sets into one data set. Simple examples are where a road network is overlaid onto cadastral boundaries, or

where neighbouring cadastral maps are joined together, possibly after one is rescaled.

As mentioned in Section 4.2.2, the Planner Broker decides upon the order and type of processing required for each data set. The data set produced by the Transform Broker depends upon the recommended operations listed by the Planner Broker. From Pascoe & Penny's (1995) notation, the decoding, translating, and encoding of S_* data set(s) from n Data Agencies, produces temporary data sets T . In addition p temporary data sets can be merged to produce another temporary data set T_j which is denoted by: $\{S_i:1 \leq i \leq p\} \xrightarrow{\text{merge}} T_j$

An alternative notation can be used to describe the merging of two or more data sets. This notation is:

$$\begin{array}{l} T_i \xrightarrow{*} T_j \\ \dots \\ T_m \xrightarrow{*} T_n \end{array} \left\| \begin{array}{l} \text{merge} \\ \rightarrow T_p \end{array} \right. \quad \text{or} \quad T_i \left\|_{i=1}^n \begin{array}{l} \text{merge} \\ \rightarrow T_p \end{array} \right.$$

This alternative notation is useful when data sets prior to merging undergo different transformations that are to be denoted. For example:

$$\begin{array}{l} T_i \xrightarrow{\text{decode}} T_j \xrightarrow{\text{translate}} T_k \\ \dots \\ T_m \xrightarrow{\text{translate}} T_n \end{array} \left\| \begin{array}{l} \text{merge} \\ \rightarrow T_p \end{array} \right.$$

6 Examples

Data integration at a National level would involve multiple user requests. This is because user needs vary, and is influenced by the GIS application being developed. In this regard, the following are some possible examples which may exist at a National level of data integration:

6.1 Single data set request

Assuming that a Data Agency formulates a query with multiple criteria, however, the Select Broker reports back with only one data set or alternatively, the user maybe only be interested with one data set. Assuming that the selected data set is in GINA format and the requested destination data set is to be in Arc/Info format.



At the Data Center, the Selector Broker would process the query and report back to the user the single metadata corresponding to the GINA format data set (assuming that the data is available). There will be no need for the Selector Broker to rank the metadata because there is only one acceptable data set. The user then selects the metadata which is then passed to the Planner Broker. The Planner Broker would decide upon the operator(s) needed to be performed on the GINA data set based upon the information obtained from the metadata. Applying the notation described in Section 5, the following processing is needed:

$$Q \xrightarrow{\text{select}} \mu \xrightarrow{\text{plan}} \tau(S_{GINA}) \xrightarrow{*} D_{ARC/INFO}$$

The selected GINA data would then be sent to the Transform Broker. The single data set is operated upon based upon the processing recommendations stated by the Planner Broker. Assuming that the data set can be decoded, translated, and encoded in that order. The operators are selected and the transformation is done. Once completed, the data is moved to the user Data Agency. The overall transformation is denoted by:

$$S_{GINA} \xrightarrow{\text{decode}} T_1 \xrightarrow{\text{translate}} T_2 \xrightarrow{\text{encode}} T_3 \xrightarrow{\text{move}} D_{ARC/INFO}$$

where $T_1 \dots T_3$ are temporary data sets. The entire transformation is abbreviated to: $S_{GINA} \xrightarrow{*} D_{ARC/INFO}$

6.2 Multiple data set request

A user Data Agency formulates a query with multiple criteria. The Select Broker performs the search and reports back with a listing of metadata satisfying the criteria. At the Data Center, the Selector Broker would process the query and report back to the user with a prioritized listing of all data sets satisfying the criteria. Assuming that the user then selects more than one data set, for example four data sets which are, two data sets with Arc/Info format, one in DXF format, and the other with a GINA format.

Using the selected metadata, this is then passed to the Planner Broker. The Planner Broker would then order the data sets in terms of the required operations

needed to be performed on each data set. Applying the notation described in Section 5 the following represents the overall processes followed:

$$Q \xrightarrow{\text{select}} \{\mu_i: 1 \leq i \leq 4\} \xrightarrow{\text{plan}} \tau(S_i: 1 \leq i \leq 4) \xrightarrow{*} D$$

The selected data sets $\{S_i: 1 \leq i \leq 4\}$ would then be copied by the Transform Broker. Based upon the operator recommendations stated in the Planner Broker the data sets are either decoded, translated, merged, or encoded. If no data merge is stated by the user Data Agency then the data sets are kept separate, however, if data merges are required then the Planner Broker will guide the Transform Broker as to which data sets can be merged. The final operation executed by the Transform Broker is the move operation which transports the data sets to the user Data Agency.

Assuming that the Planner Broker recommends to the Transform Broker the following:

- The two Arc/Info data sets are denoted by S_{AI-1} and S_{AI-2} . They require decoding, translating, merging and moving operations.
- The DXF data set is denoted by S_{DXF} , require decoding, translating, encoding and moving operations.
- The GINA data set is denoted by S_{GINA} , require decoding, translating and moving operations.

Using the notation which defines these operators, the following representation is obtained:

$$\begin{array}{l} S_{AI-2} \xrightarrow{\text{decode}} T_1 \xrightarrow{\text{translate}} T_2 \xrightarrow{\text{merge}} T_5 \xrightarrow{\text{move}} D_{AI-1} \\ S_{AI-2} \xrightarrow{\text{decode}} T_3 \xrightarrow{\text{translate}} T_4 \xrightarrow{\text{merge}} T_5 \xrightarrow{\text{move}} D_{AI-1} \\ S_{DXF} \xrightarrow{\text{decode}} T_6 \xrightarrow{\text{translate}} T_7 \xrightarrow{\text{encode}} T_8 \xrightarrow{\text{move}} D_{AI-2} \\ S_{GINA} \xrightarrow{\text{decode}} T_9 \xrightarrow{\text{translate}} T_{10} \xrightarrow{\text{move}} D_{AI-3} \end{array}$$

where $\{T_1 \dots T_{10}\}$ are temporary data sets and D_{AI-1} , D_{AI-2} , and D_{AI-3} represent the final Arc/Info, destination data sets. This notation is abbreviated to:

$$S_* \xrightarrow{*} D_*$$



7 Summary

In this paper the concept of data integration was subdivided into multiple processes which is implemented by three Brokers called the Selector Broker, the Planner Broker, and the Transform Broker. The role of these Brokers were defined, reflecting the nature of their data processing. These processes all work in synergy to facilitate integration of data sets from distributed databases. When put together, the concept of a virtual Data Center operating at the National level is established which is made accessible to multiple Data Agencies.

In presenting the model, Pascoe & Penny's (1995) data transfer notation was extended to include specific operators executed by each Broker. The new operators are: the Select Operator ; the Plan Operator ; and the Merge Operator. Using these new operators and Pascoe & Penny's (1995) operators, the retrieval and processing of data sets, starting from user queries to final data destination(s) is denoted.

The conceptual model presented a working model for further researches in each of the three listed Brokers and the operators.

References

- Abel D. J., B.C. Ooi, K. & Tan, S. (1998), 'Towards integrated geographical information processing', *International Journal for Geographic Information Systems* Vol: 12(No: 4), pp: 353-371.
- Aronoff, S. (1993), *Geographic Information Systems : A Management Perspective*, WDL Publications, Ottawa, Canada.
- Ballou, D. & Pazer, H. L. (1985), 'Modelling data and process quality in multi-input, multi-output information systems', *Management Science* Vol: 31(No: 2), pp: 150-162.
- Denness, I. (1997), 'Data for local authorities. part 1: Discovering what's out there', *Mapping Awareness* 10(8), pp: 31-33.
- Dueker, K. J. & Vrana, R. (1995), 'Systems integration: A reason and a means for data sharing', *Sharing Geographic Information* pp. 149-155. Editors: H. J. Onsrud and G. Rushton.
- Flowerdew, R. (1991), 'Spatial data exchange and standardization', *Geographical Information Systems, Principals and Applications* 1, pp: 357-387. Editors: Maguire D. J., Goodchild M.F. and D.W. Rhind. Goodchild, M. F. (1995), 'Sharing imperfect data', *Sharing Geographic Information* pp. 413-425. Editors: H. J. Onsrud and G. Rushton.
- Green D. R., D. R. & Corbin, C., eds (1996), *The Integration and Interoperability Challenge - the Need for OpenGIS, Vol. 1*, Association for geographic information (agi) source book.
- Johnson B. D., E.P. Shelly, T. M. M. & Callahan, S. (1991), 'The FINDAR directory system: A meta-model for metadata', *Metadata in the Geosciences* pp. 123-137. Editors: D. Medyckyj-Scott, I. Newman, C. Ruggles and D. Walker.
- Kevany, M. J. (1995), 'A proposed structure for observing data sharing', *Sharing Geographic Information* pp. 76- 100. Editors: H. J. Onsrud and G. Rushton.
- Marr A. J., R. T. Pascoe, G. L. B. & Mann, S. (1998), 'Development of a generic system for modelling spatial processes', *Computers, Environment, and Urban Systems* Vol: 22(No: 1). Published by: Elsevier Science Ltd., Great Britain.
- Pascoe, R. (1996), 'Data sharing using x.500 directory', *First International Conference on Geoprocessing* pp. 689- 698. Spatial Information Research Center (SIRC), University of Otago, New Zealand.
- Pascoe, R. & Penny, J. (1990), 'Construction of interfaces for the exchange of geographic data', *International Journal of Geographical Information Systems* 4(2), pp: 147-156.
- Pascoe, R. T. & Penny, J. (1995), 'Constructing interfaces between (and within) geographical information systems.', *International Journal of Geographic Information Systems* 9(3), pp: 275-291.
- Peel, R. (1997), 'Unifying spatial information', *Mapping Awareness* 10(8), pp: 28-30.
- Tayi, G. & Ballou, D. P. (1998), 'Examining data quality', *Communications of the ACM* Vol: 41(No: 2), pp: 54-57.
- Walker D.R.F, Ruggles C.L.N., N. A. & Medyckyj-Scott, D. (1992), 'A system for identifying datasets for GIS users', *International Journal for Geographical Information Systems* 6(6), pp: 511-527.
- Wang, R. & Strong, D. (1996), 'Beyond accuracy : What data quality means to data consumers', *Management Information Systems* Vol: 12(No: 4), pp: 5-34.