

Design considerations for minimising the inappropriate use of spatial data in a GIS

Greg M. Byrom¹, Dr. Richard T. Pascoe²

¹Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
Phone: +64 3 479-7611 Fax: +64 3 479-7586
Email: byromgm@imager.otago.ac.nz

²Department of Information Science
University of Otago, Dunedin, New Zealand
Phone: +64 3 479 8321 Fax: +64 3 479 8311
Email: rpascoe@infoscience.otago.ac.nz

Presented at SIRC 99 – The 11th Annual Colloquium of the Spatial Information Research Centre
University of Otago, Dunedin, New Zealand
December 13-15th 1999

ABSTRACT

The widespread use of spatial data sets from many different sources is the cause of many costly, inaccurate and time-consuming data processing problems. An organisation may spend large amounts of time and money on a spatial information project, only to find that the end result is not what was required, or even that the problem cannot be solved because the available data is insufficient. It would be desirable to have prior knowledge of the capabilities of the available data, and to analyse this knowledge in order to minimise the inappropriate use of data, thereby minimising this situation. Unfortunately, due to the widespread dissemination and use of data from different sources, the people who have this knowledge are not usually the people who are asked to do the data processing. Constraints of time and resources often force managers to require that a data set should simply be used so that a job can be completed, without appropriate consideration of the data lineage. This problem is more pronounced today given the increasing use of the Internet as a data source and the fact that so much of the data available through this source is undocumented and unsupported. A system that allows the layman to understand the consequences of using a data set would solve this problem. This paper discusses the development of such a system.

Keywords and Phrases: accuracy, precision, consistency, inappropriate, suitable, metadata, object-oriented.

1.0 INTRODUCTION

When an analysis involving one or more data sets is performed in a Geographic Information System (GIS), the result often is not formulated to include an indication of the amount of error it contains, even despite the fact that the input data sets contain errors. Human nature often dictates that we take analyses performed by a computer as gospel. Often the worst offender is the layperson, who simply wants a result from an analysis, and may not understand the need for knowledge about accuracy, completeness or many other factors that might affect the result. Ehlschlaeger & Goodchild (1994) state: "Even with simple applications, knowledge about the accuracy of data is often lost among the different people working on different aspects of spatial applications. How many GIS practitioners know how the errors in a [Unites States Geological Survey] digital elevation model affect their applications?"

This problem can lead to erroneous results and even mistakes in analyses, which can be costly for

organisations that rely on quality data. It is desirable to inform the layperson about the quality of the result that they obtain, to ensure that these mistakes do not occur. In this paper, the goal is to present techniques for making data users aware of their potential misuse of data, thereby enabling the user to see what effect the use of a data set will have on their analysis, in order to minimise any inappropriate use of the data. “Inappropriate use” is defined here as being the use of data in an application for which the data is unsuitable.

The following section explains this goal in greater detail, and gives some examples of problems that might occur in the real world that could be solved by achieving the goal. Section 3.0 analyses some existing research to show what approaches have been adopted to solve this problem, and the advantages and disadvantages of these approaches. Two new approaches to achieving the goal are introduced and described in Section 3.2. These are then compared and evaluated using the criteria listed in Section 3.6.1.

Section 4.0 describes the design of experimental systems used to evaluate the two approaches proposed in Section 3.2. The designs are discussed and evaluated against criteria for a good design. In Section 4.4, rules for the appropriate use of data are developed from the criteria in Section 3.6.1, and a proposal for a hybrid of the two approaches is introduced and discussed.

Section 5.0 deals with future research – the implementation of the approaches, and specific and general outcomes that are expected from such an implementation.

The conclusions from this research are given in Section 6.0.

2.0 THE GOAL

Inappropriate use of data often occurs when data is used in a way that was not intended when it was captured. This is because the intended use of a data set often determines the characteristics of the data set, such as its scale and accuracy (in the case of spatial data sets), amongst other characteristics. For example, say a user makes a query to match cadastral parcels with box numbers along the streets in a city. The system forwards the query on to a *Selector Broker* (Ramroop & Pascoe, 1998) which determines the available data sets. These are a street network data set containing box numbers at a scale of 1:40000, and a cadastral parcel data set of the same area captured for a different purpose at a scale of 1:20000. However, we may discover that although the cadastral parcel data set is fine for our purposes, the difference in scale means that the street network data set is unsuitable, so the box numbers cannot be matched accurately with cadastral parcels. Consequently, the use of the street network data set for this purpose is inappropriate. The user needs to be informed about this problem before using the street network data set and producing an erroneous result.

As another example, imagine a Digital Elevation Model (DEM) that is used by an engineer for an analysis of surface water runoff to determine the positions that would be most suitable for placement of stormwater drains. The error field of the DEM (Ehlschaeger & Goodchild 1994) could be used to determine the accuracy of this analysis, and the engineer could then decide whether this was acceptable. The engineer would need to analyse the surface of the error field in order to discover local trends. He may discover that the data set is appropriate for some areas of the DEM but not for others because of the variation in the amount of error over the area covered by the DEM.

Now consider that the DEM might not have been captured for the purpose of determining storm water catchments, but for a project that required much less accuracy, e.g. an analysis of sea level change due to global warming. Sharing this data set with an engineer for this purpose might not be an appropriate use of the data set due to the difference between the accuracy requirements for determining stormwater catchments and the accuracy requirements for an analysis of sea level change.

The goal of this research is to minimise these types of problems. This is done by informing the user of the consequences of using a particular data set for an analysis, and allowing the user to decide based on the information provided to them whether to go ahead and use the data set or not. Of course, any system that is developed for this purpose cannot hope to completely eliminate all inappropriate data use. It is impossible to detect inappropriate data use due to a data set being flawed, such as a data set that contains errors, omissions or false data – the system cannot reasonably be expected to know about these errors and therefore cannot warn the user that the data set is inappropriate. However, the system *can* detect inappropriate data use due to a data set having metadata that shows it is inadequate for the analysis.

3.0 ACHIEVING THE GOAL

3.1 Existing Research

Much of the existing research in this area centres around data quality and selection of data sets based on suitability of use (Ramroop & Pascoe 1998, Dutton 1996, Wand & Wang 1996, Allan 1994, Goodchild – various). Much has also been done on error in spatial databases (Hunter & Beard, 1992). However, little research has been done on ensuring that end users use data sets appropriately. Cracknell (1998) attempts to make the Remote Sensing community aware of this problem, but does not fully address it: “There is no simple answer to the question ‘what exactly gives rise to the signal detected and recorded in a pixel in a remotely sensed image?’ The main point to be made is to try to ensure that it is realised that there *is* a problem and to give some indication of the nature of that problem”. (Cracknell, 1998).

A *Selector Broker* could be used to determine the available data sets. A Selector Broker is an entity that selects data sets for use based on the user’s requirements, and arranges them in order of the user’s priority (Ramroop & Pascoe, 1998). This paper extends the Selector Broker concept to enable the use of any data set, by allowing the user to see what effect the use of the data set would have on the result, and decide whether this is acceptable. Essentially the attributes (metadata) of the data set need to be compared with the user’s requirements in order to establish the suitability of the data.

As we have seen in Section 2.0, the problem of use of unsuitable data sets often stems from the use of data sets for purposes for which they were not originally intended. This is a common occurrence, as Allan (1994, page 218) indicates: “...the segmentation based on hydrological units rarely coincides with administrative or natural boundaries; the public utilities, public transport and other infrastructures are rarely spatially organised according to consistent concepts; and the same is true of health services, the police, census units, electoral units etc”.

The inappropriate use of spatial data could be overcome by storing data quality information with the data set itself, such as storing locations in the form of quadtrees that also specify scale and accuracy as proposed by Dutton (1996). These quadtrees are a form of metadata, but are not separated from the data itself, so the user cannot get a result that does *not* include data quality information. This ensures that the user is aware of the limitations of the data set. Ehlschlaeger and Goodchild (1994) state that tools for keeping this data quality information up to date are already in development: “Ongoing research at NCGIA is creating a suite of public domain software tools that explicitly define error in maps, propagate map errors through all spatial analyses, and visualize the implications of map error on any and all spatial analyses”.

An important aspect to the problem of inappropriate use of data is defining what data quality information is to be stored. A good definition of data quality is given by Wand & Wang (1996, Page 87). The same work also describes several ways (*data quality dimensions*) in which data sets may be considered “suitable” (or otherwise) for use in a given situation (Wand & Wang 1996, Pages 93-94).

3.2 Approaches to the problem

A practical and workable approach should decide how appropriate a data set is for use in any situation. Such decisions are very much dependant upon the requirements of the user, which can vary widely. These requirements can be expressed in many different forms, such as in the form of a query which must be analysed by the system to determine which data sets meet the requirements.

Two approaches are explored in this paper:

- An object-oriented approach, where the data is encapsulated and its use is limited by a set of operations;
- A metadata approach, where the data set’s metadata is compared with the requirements of the analysis to determine whether the data set is appropriate for the analysis.

The objective of this research is to prototype these two approaches, and to evaluate their effectiveness for dealing with simple problems such as the street network and cadastral parcel problem above. The effectiveness of an approach can be evaluated in terms of the extent to which a user can be informed of the implications of their use of a particular data set. This information can be provided to the user in the form of a document file describing the effects of using a data set on their analysis, or simply in the form of a dialog box presented to the user when opening a data set for use.

3.3 Criteria for use of data sets

Although the requirements of particular users may vary widely, there are always certain criteria that need to be matched and a comparison made between what the user requires and what a particular data set can supply. Wand & Wang (1996) refer to these criteria as *data quality dimensions*, and give a good general definition of each.

Some criteria for the appropriate use of data sets are:

- Accuracy - differences between the view perceived directly from the real world and the view perceived through the information system.
- Precision - the smallest possible difference between the real world and its representation in an information system.
- Completeness - the amount of data included in the data set.
- Content - the type of data specified by the data set.
- Scale – the relationship between the size of features in the data set and their equivalent features in the real world.
- Timeliness - the temporal difference between the data set and the real world that is represented by it.
- Consistency - the uniformity of representation of features between data sets used in the analysis.
- Other criteria used to determine the suitability or otherwise of a data set might be the data file size, ownership, format and availability of the data, and the characteristics of the application itself.

3.4 An Object-Oriented (OO) approach

In this approach a set of operations that can be performed on the data set is encapsulated with the data set. The use of a data set involves one or more of these operations. Some of the operations are available for all data sets; others are specific to a particular data set. If a particular operation generates a warning that the data set may be being used inappropriately, the user may choose to ignore the warning if they wish. Any use of the data set that involves operations that are not available for the data set, is defined as inappropriate, unless the user overrides this. Any use of the data set that involves only those operations that are encapsulated with the data set would be defined as appropriate.

For example, a street network data set at a scale of 1:40000 might have an operation labelled “overlay with no better than 1:40000 scale data set”. If the user requires that this data set be overlaid with a cadastral parcel data set at a scale of 1:20000, they will find that this operation is not possible since it is not an operation that is encapsulated with the street network data set. This might be the case if the data set were not originally intended to be used for such an analysis because it requires a result that the data set is not capable of providing.

3.5 A Metadata approach

The spatial and temporal resolution of the data and its spatial domain, temporal period, ownership and copyright issues, accuracy, format and location are examples of metadata that might be used in this approach, because they are all factors that affect the suitability of data for use in various ways. The importance of storing this metadata and keeping it up to date is reinforced by Ehlschaeger and Goodchild (1994): “Unless map construction metadata exists, issues such as product uncertainty and sensitivity, risk analysis and others are conceptually difficult”.

The user’s requirements can also be called metadata. For example, a user might make a query that requires a street network data set at a scale of 1:20000. If the only available street data set has a scale of 1:40000, as determined by the data set’s metadata, then its use would be inappropriate.

The matching of these two sets of metadata allows us to formulate a set of rules for the use of the data set. This set of rules might specify that the data cannot be used at all, or it might specify some constraints on the use of the data to ensure that the result obtained falls within the limits of what the user requires.

3.6 Comparing the approaches

3.6.1 Criteria for a good approach

In order to make data users aware of their potential misuse of data, a good approach should:

- a) Perform as few comparisons as possible between the data set and the requirements of the analysis. This makes the system efficient, ensuring that users do not get frustrated with having to wait for the comparison to be made, and therefore they are less likely to simply “skip this step”.
- b) Be capable of modifying the data set’s metadata and/or set of possible operations to reflect modifications that are made to the data set itself. This produces an entirely new data set with a new set of operations and/or metadata. This must be done to ensure that new data sets are always compatible with the system.
- c) Be able to be integrated into an existing GIS, so that users can determine the inappropriate use of spatial data in their own applications. This ensures that it is always possible to determine inappropriate use, no matter what type of GIS is being used.
- d) Have good functionality. This includes:
 - Showing the user what is wrong with the way in which they are using a data set.
 - Showing various options for how it can be used instead.
 - Allowing the user to use the data set anyway, if they wish.
- e) Interact as little as possible with the user, except to inform the user about the data set’s capabilities and the way in which it is being used. This is to ensure that the user does not have to make technical decisions about the use of data sets – the system should be capable of making these decisions for them. The user should not need to know how to make a data set compatible with the system.

It is not considered necessary for the system to be able to modify legacy data sets for use with one of the approaches. Although it is possible to generate suitable metadata and/or a set of possible operations for a legacy data set, this operation is beyond the scope of the system. Instead, the system requires a special type of “compatible” data set, but this should not preclude the use of the data set in a normal situation *without* the use of a GIS that has been designed to minimise inappropriate use.

3.6.2 Discussion

The two approaches differ in the type of information that is stored about the data set and the way in which it is used. The Metadata approach requires only that metadata be stored, and a comparison is performed on this metadata that allows us to decide whether the data set will be appropriate or not. The OO approach specifies only certain operations that are available for the data set, so that we do not have to *decide* whether the data set is appropriate or not. If the required operations are not supported for a particular data set, then it is deemed inappropriate and cannot be used for that purpose, although the user can override this.

Structuring our comparison upon the above criteria we find that:

- a) The metadata approach requires that a comparison between the metadata and the requirements of the analysis must be performed for each aspect of the stored metadata – comparing accuracy, precision, completeness etc. separately. The OO approach only requires a single comparison of available operations (for the data set) and required operations (for the analysis) to be made.
- b) Both approaches allow for the modification of a data set and creation of a new metadata set, which can then be used again by the same system. The OO approach requires us to generate a whole new set of possible operations in order to reuse the data set. This makes the OO approach a little more difficult to implement than the Metadata approach if it is to completely fulfil our requirements.
- c) The OO approach is active while the data set is being used, and therefore is difficult to integrate into an existing GIS because it needs to continuously interact with the GIS in order to determine inappropriate use. The Metadata approach can act *before* the data set is used, at the time when data sets are selected for use. This makes integration simpler because it can be performed by a separate utility and does not have to be a part of the GIS itself.
- d) The Metadata approach is more capable of showing the user what is wrong with the way in which the data set is to be used, but this information is presented to the user at the time when the data set

is selected for use, which may not be helpful. The OO approach presents the user with information about the suitability of the data set at the time when it is actually used. This is more useful but the OO approach is a little restrictive; since it only allows certain operations to be performed, it cannot tell the user what is wrong with an operation that is *not* allowed to be performed.

- e) Both approaches require little or no user interaction.

4.0 EXPERIMENT DESIGN

The design of a system to minimise the inappropriate use of spatial data in a GIS must focus on the criteria discussed in Section 3.6.1. These criteria in turn are based on the goal of presenting techniques for making data users aware of their potential misuse of data, thereby ensuring that analyses are not simply performed on inaccurate and incomplete data sets giving an erroneous result.

Two prototype systems for enabling the user to see the effect of their use of a data set are in development. One of these systems deals with the OO approach, and the other deals with the Metadata approach. The system compares the data set's metadata and/or set of possible operations with data about the requirements of the analysis being performed, when the data set is opened for use.

Both systems follow a similar set of rules for deciding on the appropriate use of a data set. The most important rules to be used for this purpose are defined in Section 4.4. The systems consist of an application that interacts with the data set and with the user's query to match the possible uses for the data set with the required uses, in order to determine whether the data set is being used appropriately.

4.1 System limitations

Ideally, a system for minimising the inappropriate use of spatial data should be able to deal with any type of data set. However there are limitations on the data sets that can be used with the system, since the data sets must either have suitable metadata and/or suitable operations must be encapsulated, and the system must be able to use these to determine how appropriately the data set is being used.

For the Metadata approach, the ability to create data sets that are compatible with the system is dependent upon the storage of suitable metadata with the data set. This metadata is stored during capture and updated during modification of the data set, to ensure that all the required metadata is available. The metadata that is stored would reflect the data quality dimensions given by Wand & Wang (1996).

For the OO approach, suitable operations are encapsulated that reflect the capabilities of the data set. These operations are determined at the time the data set is created or modified. The actual operations that are used depend on the intended use of the data set and the information that can be extracted from it. Some of these operations could be inferred from the data quality dimensions given by Wand & Wang (1996), others would be specified by the data agency that captures the data set.

All relevant data quality dimensions should be catered for by the system. The particular data quality dimensions that are relevant would depend on the type of inappropriate use that is being minimised. Ideally every kind of inappropriate use would be able to be determined, but it is more realistic to only deal with those types that are easily determined from the available metadata and from the limitations imposed by the capture process.

Analyses that are performed by the user will have many requirements, and no system can deal with all of the possible requirements of every analysis. The system can only be expected to deal with basic requirements such as scale, accuracy, coverage, feature set, and so on. It should be noted that inappropriate use of data would still be possible if a user performs a particular analysis that has requirements that are beyond the abilities of the system.

4.2 Prototype limitations

The prototype systems that are in development have three main limitations. These are as follows:

- Data set processing is not catered for in the prototype, therefore it is not possible to ensure that, for example, data sets that are scaled to create a new data set at a different scale contain the new information needed to determine inappropriate use.
- The prototype cannot be integrated into an existing GIS. Although the principles explored in the

prototype can be translated to work in a fully functional GIS, the prototype is to be used on its own to demonstrate the practicality of the system.

- The metadata prototype deals only with a limited set of metadata and can only determine certain types of inappropriate use that can be inferred from this metadata. The OO prototype deals only with a limited set of possible operations, and can only determine certain types of inappropriate use from these.

4.3 System requirements

Each system should meet the following requirements:

- The system must be as transparent as possible to the user. The user should not need to know how to make a data set work with the system. Making data sets compatible with the system (generation of metadata and/or possible operations) should be performed at the time of capture by the data agency responsible for capture. The system itself is then responsible for keeping these metadata and/or operations up to date.
- These modifications must be minimal and should not in any way affect the application of the data set to other problems by other data agencies.
- If the data set is used in a GIS application, this GIS application should not have to be modified extensively by the user to incorporate features for the system. The system should rely as much as possible on functionality built into the data set itself, rather than the GIS application, to determine whether the data set is being used appropriately.
- The system must inform the user of the potential for erroneous or inaccurate results as a result of their use of a data set for a particular purpose. It should then give the user the opportunity of correcting this, by either not using the data set at all, or using it in a different manner, or modifying their requirements. It should also allow the user to use the data set anyway if they choose to do so, rather than preventing its use altogether.

4.4 Rules for the appropriate use of data

Whichever approach is used, the rules for using data sets appropriately remain the same. The set of rules that is determined from the set of possible operations in the OO approach should be the same as the set of rules that is determined by the metadata in the Metadata approach.

Some of the more useful rules for the appropriate use of data are:

- Difference in scale from other data sets used should not be large (just how much is “too much” would be determined by the data sets involved and the application)
- Feature set should include the required features, as determined by the application
- The data set should cover the correct area completely
- The data set should have the required accuracy over the area of interest
- The data set should have the required precision. If the data set is in a raster form, the resolution should be within certain limits. If the data set is in a vector form, coordinates should be stored to the required number of significant figures. (Again, the other data sets used and the application would determine these requirements).
- The data set should not be out-of-date for the purpose for which it is being used. Timeliness would be determined by the application.
- The data set’s error field should not be too large (this is related to accuracy and precision above).
- ...and many others. The number of possible rules is too large to be listed here and any system that is developed can only cater for so many of them, so only those that are deemed to be most important will be concentrated on. Many rules could be based on the data quality dimensions defined by Wand & Wang (1996, pages 93-94).

As many of these rules as possible would need to be catered for. In the case of the OO approach, this means that the set of possible operations that is encapsulated with the data set would need to reflect this set of rules. In the case of the Metadata approach, this means that the metadata that is stored with the data set would need to cover all aspects of each of the rules – such as the age of the data set, its resolution and scale, error field, and so on.

4.5 System design for the Object-Oriented approach

Since the set of operations that are possible for each data set is encapsulated with the data set itself, all of the information that is necessary to determine the data set's suitability is in a single package. This approach is efficient because the possible operations that can be used with the data set do not need to be generated on the spot. Access to the data set is only possible through the defined operations, which act as a "wrapper" for the data set, ensuring that it is used only in an appropriate fashion. It is also necessary to have a "user override" for this approach; this can be achieved by making it an operation that is defined for every data set.

Operations that are available for a particular data set might include:

- "Overlay with higher resolution data set"
- "Overlay with newer data set"
- "Combine with data set not more than 1 year older"
- "Extract parcel ownership information from attribute database"
- "Analyse soil types to nearest 100m resolution"
- "Convert to 20m resolution or lower raster data set"
- "Reduce radiometric resolution"
- ...and many others.

Individual requirements and the information content of the data set determine the actual operations that are encapsulated with a particular data set. Different data sets require different types of operations to be stored; for example a raster data set might specify operations to do with spatial and radiometric enhancement, and a vector data set might contain operations for overlaying with other data sets, extracting feature information, and so on.

4.6 System design for the Metadata approach

In this design the system is a separate entity, such as an application that performs a comparison between the data set's metadata and the requirements of the user. This method is less efficient than the OO approach but requires less interaction with the GIS application that is being used for analysis. The system performs the comparison by matching criteria such as resolution and scale requirements, spatial extent, etc., between the data set's metadata and the analysis that is being performed by the user. The system can use either a "fuzzy logic" sub-approach or a "clearly defined rules" sub-approach in which metadata attributes such as scale and coverage are simply compared to produce a yes-or-no answer.

Rules for using the data set are generated from the metadata, and these rules are applied to a data set to check whether using the data set in a particular instance would break any of them. If any rules are broken, the data set is inappropriate for the current analysis, but the user can be informed that the data set might be appropriate if they are willing to modify their requirements, e.g. by accepting a less accurate result. These rules are based on those in Section 4.4, but are specific to an individual data set.

The types of metadata that would be useful in this approach might include:

- Temporal domain of the data set (its age).
- Temporal period (how much time it represents).
- Temporal resolution (the temporal accuracy of the data set).
- Spatial domain (*where* in space is represented by the data set).
- Spatial period (*how much* of space is represented).
- Spatial resolution (the spatial accuracy of the data set).
- Feature set (what features the data set represents).
- And many others.

4.7 Comparing the designs

4.7.1 Criteria for a good design

A good design should:

- a) Require minimal modification of any GIS application in which it is used. GIS applications that have already been built should be able to be modified easily to work with the system; the user should not have to build a specific application in order for the system to perform the required comparisons between the data set and the user's requirements.
- b) Require minimal modification of the data set in order for the approach to work with the data set. A specific data type or format should not be necessary. For example, a separate metadata file can be used to store the data set's metadata, provided this file is always kept with the data set itself.
- c) Make it easy for the user to see the problems that are detected in their use of the data set. This can be in the form of a dialog box presented to the user with a description of what effect the data set will have on the analysis and various options for its use.
- d) Be as transparent as possible to the user, to allow for use by people who know nothing about the system. The user should not have to "feed the data set into the system".
- e) Cater adequately for the various rules for the appropriate use of data, some of which are specified in Section 4.4.

4.7.2 Discussion

The OO and Metadata approaches differ greatly in design while achieving the same goal. The criteria laid out in Section 3.6.1 are reflected in the criteria for a good design, given in Section 4.7.1. These criteria are more practical in nature; they reflect how closely the *design* of an approach can be made to achieve the goal. None of the criteria can ever be completely fulfilled, and one design might achieve the goal more comprehensively than the other. However, the two approaches fulfil different criteria to a different degree, and it is expected that the best approach would be a hybrid that incorporates the best aspects of both approaches. This hybrid system is discussed in Section 5.2.

Comparing the above criteria we find that:

- a) The OO approach is difficult to integrate into an existing GIS because it needs to continuously interact with the GIS. The system that implements the OO approach needs to be activated when the data set is used, to make the necessary comparisons. In the case of the Metadata approach, the application must be capable of passing the user's metadata requirements to the system, which makes the required comparison. Therefore both approaches require some interaction with the GIS application; neither is ideal.
- b) The Metadata approach requires almost no modification of the data set itself, since the metadata can be stored in a separate data file. The OO approach is more complex because it requires that the possible operations for the data set are encapsulated with the data set, and the data set can only be analysed through the operations that are defined for it.
- c) The OO approach gives a definite yes-or-no answer as to whether a data set can or cannot be used. If a data set fails the test for inappropriate use, it is difficult to inform the user about *why* it failed and easier to simply inform them that it *did* fail and show the operations that it failed with. This is not always helpful, especially if the operations that caused the failure are technical in nature. The Metadata approach is less restrictive and can give a summary of the characteristics of data sets and what causes them to pass or fail the test.
- d) Both systems are highly transparent to the user although the OO approach is slightly ahead in this regard because it restricts the user to only certain operations, so the user does not have to intervene to make decisions about operations that are *not* possible.
- e) The OO approach caters for the rules for the appropriate use of data (Section 4.4) slightly better than the Metadata approach since the rules can be equivalent to the operations that are defined for the data set. In the Metadata approach these rules must be generated on the spot, so their definition is less clear.

5.0 FUTURE RESEARCH

5.1 System Implementation

The system is to be implemented as a prototype application that performs the functions that are required to inform the user about the suitability or otherwise of a data set. For each approach, code will be written to do two separate operations. The main application that is incorporated with the data set will control the comparison of data sets with the requirements of the analysis being performed, and updating the metadata and/or set of possible operations when the data set is processed. A separate application will make data sets compatible with the first application, by ensuring that the user supplies appropriate metadata during data capture. However this should still allow the data set to be used without the system if another data agency wishes to do so.

5.2 An OO / Metadata Hybrid?

In a hybrid approach the data set is encapsulated within the data description itself. The set of possible operations that can be performed on the data set is determined by this data description; this set is matched with the set of operations that the user requires to determine whether the data set is appropriate for a specific use.

When the user performs an analysis using the data set, the set of required operations will be 'fed into' the data description and the outcome determined, in order to see if those operations can be performed on the data set. This enables the user to see how the data set can and cannot be used, and therefore prevents inappropriate use.

6.0 CONCLUSIONS

Two methods for determining the appropriateness of data sets for use in answering user's queries have been proposed, in order to inform the uninformed user about the consequences of using data sets for their purposes. A comparison is performed between the requirements of the analysis being performed by the user and the data set's attributes. The results of this comparison are presented to the user to enable them to see what effect the use of a data set will have on their analysis. In this way the inappropriate use of spatial data can be minimised, and therefore the costs in time, money and user frustration associated with this inappropriate use can be minimised also. Real-world problems can be solved with greater accuracy, less cost and in less time when the suitability of the data sets used is taken into account.

REFERENCES

- Allan J.A. (1994) Spatial Data: Data Types, Data Applications and Reasons for Partial Adoption and Non-integration, *Visualization in GIS*, John Wiley and Sons Ltd.
- Cracknell A.P. (1998) Synergy in remote sensing – what's in a pixel?, *International Journal of Remote Sensing* 1998 Vol.19 No.11 pp. 2025-2047
- Dutton G. (1996) Improving locational specificity of map data – a multi-resolution, metadata-driven approach and notation, *International Journal of Geographical Information Systems*, 1996 Vol.10 No.3 pp. 253-268
- Ehlschlaeger C.R. and M.F. Goodchild (1994) Uncertainty in Spatial Data: Defining, Visualizing, and Managing Data Errors, *Proceedings of GIS/LIS 1994*, NCGIA, pp. 246-253 (http://www.sbg.ac.at/geo/people/elorup/diss/lit/ehl_good_1994_2/gislis.html)
- Hunter G. and K. Beard (1992) Understanding error in Spatial Databases, *The Australian Surveyor*, 1992 Vol.37 No.2 pp. 108-119
- Ramroop S. and R. Pascoe (1998) Notation for National Integration of Geographic Data, *Proceedings of the 10th Annual Colloquium of the Spatial Information Research Centre*, University of Otago, Dunedin, New Zealand, 16-19 November 1998.
- Wand Y. and R.Y. Wang (1996) Anchoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM*, November 1996 Vol.39 No.11 pp. 86-95