B03.1

Information led or information misled?

Andy Gill, Consultant, InfoTech Enterprises Europe

Abstract

Many organisations realise the benefits of using geographic information (GI). However, whilst the tools are in place, and people are trained to use the increasing levels of GI available, is the data/information ready? There is the danger that well thought out strategies, based upon data used within a GIS, will be ineffective if the data are of uncertain quality and therefore unreliable. Data reliability is crucial. Ultimately, if you do not know where clients live, where problems occur, or where to target resources with confidence, then the full power of GIS will not be utilised and appropriate decisions not made.

Arguably, data is now more important than the technology available to use it. The value in addressing the accuracy of new and existing records should not be underestimated. The use of good, accurate, high-quality data will enhance and inform strategies, however this is only true if data is used correctly.

Spatially accurate, up-to-date, and fit-for-purpose data will provide useful information, enabling effective decisions to be made. Information led decision-making conforms to national models and statutory requirements. Success will follow when information leads to an effective solution rather than misleads and wastes resources.

Effective decision-making relies on access to good information; the question of "What is good GI?" is explored within this paper. The importance of using metadata to understand whether your data is fit-for-purpose, and fit for which purpose, will also be emphasised. Examples will be provided of how good GI has led the way in supporting appropriate decisions.

This paper outlines the common data issues GIS users should consider when beginning to work with data, drawing upon real-life experiences. Topics discussed include spatial accuracy and precision, data cleaning, data management and metadata, and promoting the benefits of good data.

1 Introduction

"Lack of information is a key part of the problem, and better information must be a key part of the solution." PAT 18 Report (Social Exclusion Unit, 2000).

The new Pan-Government Agreement pilot between the ODPM and Ordnance Survey is to be applauded, for it will expose geographic information (GI) to a wider audience than previously possible. This will provide base mapping data to agencies and authorities that have not previously had access to this type of information. There will be instances where further interest in utilising other datasets alongside the OS data will be generated. Also with joined-up government increasingly on the agenda, there will be an increased demand for data from multi-agencies or partnerships to achieve it, (much of this data will include geographic references). Inevitably there will be people never having used GIS before being exposed to GI and herein lies the danger that data will be ill-understood, under-used or completely abused.

The oft-quoted magical figure that '80% of information possesses a geographical component' implies there is an assortment of data available for use within a GIS; but how much of this is good geographic data?

The topic of data quality has been raised at previous AGI conferences. However, there is still a need to inform new and existing GIS users, of the value in ensuring their GI is:

- of high currency,
- fit-for-purpose, and
- used correctly.

Additionally, it is important to understand GI before analysing it, and presenting results. The user-friendliness of GIS can provide an illusion of good GI, when in reality the quality is unknown or the data is used incorrectly. To promote better understanding, metadata should be collated and fully maintained throughout a dataset's life cycle.

Geographic data is increasingly relied upon for many business processes and strategies (public and commercial), however sometimes there is scant regard for the quality of the data. This paper will display the key issues of concern and give consideration of how to resolve problems with data.

2 Good (and bad) GI

What is good GI?

For the purposes of this paper, vector data will be discussed, as this type of GI has the most potential to be flawed. Errors can occur with raster datasets due to incorrect image registration, or incorrect user interpretation. However, with vector data the problem is multiplied because of the attribute component of the data.

What contributes to GI being 'good'? There are two main factors here; namely accuracy and fitness-for-purpose, although a third could be included to cover the expertise of the user of the data. The accuracy of GI refers to how correct the geographic objects and associated attribute information is. Describing data as being fit-for-purpose can be quite subjective, therefore a fuller outline is provided in the next section.

Fit-for-purpose

A traditional problem with GI is that data is usually collected for one purpose only. Problems arise when this data is used for a different purpose to the one originally intended - the information can then be considered 'unfit-for-purpose'. Problems also occur where data was collected for use in one IT application, and cannot easily be used in a different one.

Hence, for data to be all encompassing 'fit-for-purpose', it shouldn't need to be reworked for use for a different purpose than originally intended, or for use in different applications.

The common theme running through all GI is the geography at which that dataset was collected. All GI concerns information related to people and places, involving a defined boundary, postal address, or more accurate grid reference. To be 'fit-for-purpose', GI requires high geographic precision in its collection. If GI is georeferenced to postcode unit level or precise grid references, it can be considered 'fit-for-all purposes'. High precision GI is highly flexible for producing aggregate statistics to any boundary, enabling better polices to be formulated, resources to be targeted more effectively, and improved monitoring of change. Some datasets are collected at lower levels of precision, due to lack of capability or lack of forethought in the collection process. Some organisations do not possess the resources (time, expertise and funding) to obtain high precision GI, or understand the benefits this can bring.

However, high precision GI cannot be assumed to be completely 'fit-for-purpose'. The quality of the information also needs to be taken into consideration – in other words the geographic objects and the attribute information behind them.

Example of fit-for-purpose data: London Borough of Croydon Community Safety Team

To effectively tackle crime and disorder at the local level, resources must be targeted to areas of need. Croydon Community Safety Team (www.croydon.gov.uk/execdept/Crime) maintains an archive of depersonalised crime incident data from the Metropolitan Police. Each crime incident is geocoded to postcode unit level, with six-figure easting and northing grid references. Their data is of such high precision, it enables hotspots to be visualised to assist with the better design and allocation of crime and disorder reduction resources. The Home Office crime reduction website (www.crimereduction.gov.uk/toolkits) advocates the analysis of crime data in this manner to identify hotspots, which can then be focussed upon to support a more informed decision-making process. Diagram 1 demonstrates the benefits of highly precise GI for Croydon Community Safety Team.

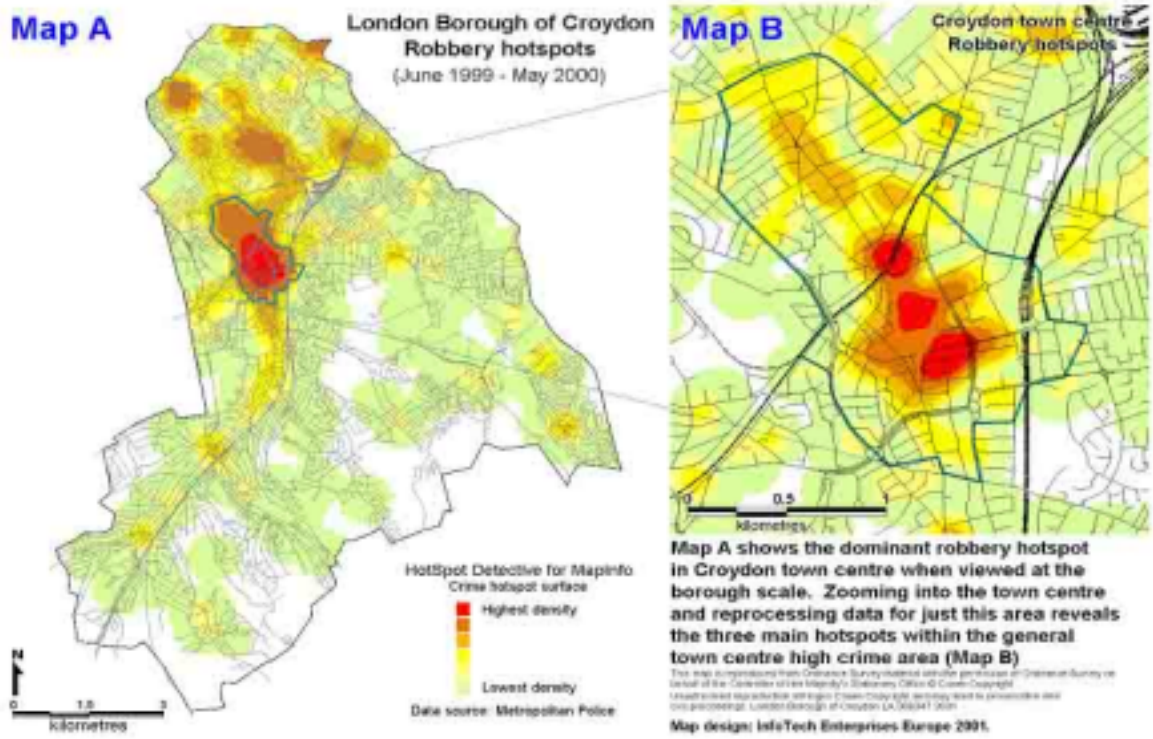

Diagram 1 – Fit-for-purpose data used in crime reduction.

3 Why Have Good GI?

GI is widespread throughout organisations, and as such will contribute to government and business strategies. Strategies will only be as good as the data used in their creation; the 'junk in, junk out' principle applies strongly. Inaccurate data or incorrect usage will lead to a strategy of limited use causing resource wastage, and loss of confidence in the decision maker. The main benefits of good GI are discussed below.

Improving trust in public services

The PIU (Performance and Innovation Unit, 2002), highlighted one barrier to sharing information, as "a lack of public trust in the way that the public sector handles personal information and the security of that information, and some concern about the risks to personal privacy from technological change". If everyone using data concerning personal information ensures it is accurate, reliable, and used correctly in compliance with the Data Protection Act, trust will be improved.

Joined-up services

Many government initiatives are concerned with improving public service delivery by 'joined-up' working. As geographic referencing is common to many datasets, this would seem to be the sensible way to join-up information. As public service delivery is reliant upon information concerning people (identity and location), it is imperative the GI is correct to enable intelligence led resource targeting.

Monitoring of services

Consistently accurate GI is needed for monitoring the effectiveness of established strategies. This is reliant on the same datasets being available throughout the project lifecycle. Data should be consistently recorded and be of comparable quality. PAT 18 (Social Exclusion Unit, 2000) identified this problem; "Lack of information has kept communities in the dark. They have not been able to make informed decisions about whether local and national government is improving neighbourhood outcomes, many of which are not immediately visible."

Increasing the applicability of GI (& recover costs for government)

The IGGI 'Principles and Practice of Sharing and Trading Government Information' guide (Intra-Governmental Group on Geographic Information, 2001), recognises "government information will need to be made more accessible to a wider audience, and that the development of value-added information products and services is something to be encouraged, possibly in co-operation with the private sector." This is an important development, as many public sector bodies have ever-decreasing funds with which to maintain a consistent level of service. The influx of funding through value-added products and services could be used to improve data collection and accuracy.

Examples of where good GI can be applied:

Example 1 - Policing National Intelligence Model (NIM)

Policing has altered from reactive investigation, towards the proactive targeting of offenders on the basis of intelligence, and identification and elimination of the causes of, and the opportunities for, offending behaviour. This problem oriented policing philosophy forms the basis of the policing NIM. Highly precise, accurate and relevant GI is necessary to form the foundation of the intelligence required to make the NIM a success.

Example 2 – PAT 18 Neighbourhood Renewal

One recommendation of the PAT 18 report (Social Exclusion Unit, 2000) was the creation of small area statistics, which would be an improvement on ward level data for identifying local problems (and addressing them).

The Neighbourhood Statistics service was created to accommodate this requirement. Currently, datasets are available to ward level, however this is not suitable for facilitating a full understanding of local problems and solutions. When the 2001 census statistics are released in 2003, they will be available as small area statistics, which will provide more detail than is currently permissible at ward level.

The dilemma facing many organisations is the need for retaining the privacy rights of individuals, whilst retaining high geographical precision. Small area statistics and postcode units may offer a way forward for the provision of data which meets both criteria.

Example 3 - Reinsurance

In the wake of September 11th, reinsurance underwriters are introducing limits to ensure that geographical areas (such as postcode or zip-code areas) do not contain too many insured properties. To effectively achieve this, accurate GI is required concerning where insured properties are located.

4 What Leads to Bad GI?

Bad data will affect all organisations at some stage, having disastrous consequences if not managed. The PAT 18 report (Social Exclusion Unit, 2000), identified four main problems that poor information has led to:

- "Lack of awareness of neighbourhood problems and trends by communities, local and national government;
- Poor diagnosis of problems has led to poor government strategies and resources allocation;
- Lack of information has forced new programmes to spend time and money collecting new information; and
- Lack of information has meant it is difficult to tell whether policies work."

The above is directed towards data used by public bodies, but can be applied to any business. The PIU report (Performance and Innovation Unit, 2002) referring to public attitude research on privacy and data use, reveals there are public concerns over data handling errors, infection with inaccurate data, and misidentification.

There are many wide ranging and varying factors leading to poor quality GI. The main issues are outlined in the following sections.

Lack of metadata

This results in a lack of understanding of **what** datasets are out there, **who** is responsible for them and **who** can have access to them, and **how** they can be used. The importance of an established metadata approach to maintaining datasets is discussed in a later section.

Lack of maintenance

Without regular maintenance, the value and currency of data declines. This is a major contributory factor to poor data quality. Whilst many owners have processes and procedures for maintenance, there are countless others not adopting the same ethos.

Lack of joined-up geography

Problems occur when agencies collect datasets according to their own geographical referencing system. For example, police data can be referenced by beat areas, which bear little or no resemblance to ward boundaries. The lack of coterminous boundaries is heightened with regular administrative boundary changes, making geographical comparisons over time difficult, (the PAT 18 report (Social Exclusion Unit, 2000), recommended that the number and frequency of boundary changes should be reduced.). One has to consider there is also a lack of joined-up attribute information. For example, some police forces record ethnicity differently to the census (and other agencies such as Youth Offending Teams), making comparisons difficult for ethnic monitoring purposes.

User incompetence/inexperience

This is linked to a lack of metadata, and training for key personnel in the basics of GI, and how to make the best use of datasets. Data collectors and inputters need to be aware of the value of their tasks. Custodians of GI should ask questions of how their data is being used, and invite others to comment on how the dataset could be bettered to improve joined-up working.

Data not given the priority it deserves

Much GI originates from data primarily collected for administrative processes, and then attached to geographic objects as an afterthought. This lack of planning can result in duplication of effort, and the data may have to be recaptured due to the original data not being fit-for-purpose. There are also many organisations where protocols to clean data are severely lacking. Some data owners are willing to accept their data is not as accurate as it could be, but reluctant to resolve the cleanliness, thereby creating a risk when other people use the data. The PAT 18 report (Social Exclusion Unit, 2000) sums this up observing, "the collection of data in many organisations is not seen as important enough at senior levels. It is rarely given priority, non-data specialists fail to appreciate its importance until it is too late, and frequently the data collectors are inadequately trained and not made aware of why the data is necessary and how it will be used."

5 How to Remedy Bad GI

There are ways to remedy many of the problems experienced, which tend to be based on improving data handling and maintenance awareness; in other words making data a priority.

Information sharing protocol

Many multi-agency partnerships are adopting information sharing protocols. The PIU report (Performance and Innovation Unit, 2002) advocates this for removing the complication of sharing data. Traditionally, many data custodians are wary of sharing data, fearing loss of control in maintenance leading to diminished

data quality. A protocol would promote awareness of who is responsible for maintenance and correction of errors.

Data quality can be improved through information sharing. "More active databases with more users increase the average number of times a single entry may be accessed, checked and validated. As such, errors may be detected and corrected sooner." (Performance and Innovation Unit, 2002). However, the PIU report mentions that data sharing can be detrimental to data quality. For example, users unrelated with the original data collection may create errors, which may be spread amongst partners, and the existence of multiple users may obscure who is the data custodian.

You will never find completely error-free GI, but with appropriate measures in place to tackle errors, information sharing can be successful. The benefits of information sharing are clear, as data custodians will not want to share their information if it is not of sufficient quality and unless guidelines are given over its use. Information sharing can therefore be regarded as a driver towards good GI.

Adopt standards

The BS7666 standard exists for defining the location of property and places, and BS8766 for defining the names of people. In the creation of metadata records, standards should be used also. However, even if data conforms to standards it can still be incorrect.

Introduce data maintenance processes & procedures

Processes and procedures should be created and adhered by to ensure continuous maintenance of data. GI quality should not be addressed once, and then assumed to be accurate forever after. GI will always be subject to inaccuracy due to the constant changes inherent in everyday life.

Organisations also face the problem of staff turnover. Established procedures are important to ensure knowledge of data maintenance remains within the organisation and doesn't leave with individuals. There should be wide communication over GI issues, with a view to addressing them. For example, Surrey Police have feedback mechanisms in place to ensure inaccuracies in the command and control gazetteer are identified, and that data custodians are informed to enable investigation and correction.

The ethos of 'always striving to improve data quality' should be adopted, as this will lead to improved service provision. Also, "good management of data entry, if applied consistently across organisations, should also deliver improvements in data quality." (Performance and Innovation Unit, 2002). A final action to consider is to undertake regular audits of data quality to ensure data currency.

Investment in training and information

Investment will be required to correct inaccurate GI, either by purchasing new datasets or expertise. Investment is needed to inform and train all those involved with GI. This is necessary throughout the whole process of data capture and input, maintenance and storage, and finally analysing the information and producing useful output. There is little benefit in building partnerships and sharing data if the GI is incorrect or used improperly.

Clean up the data

If your data is unclean, be it through a lack of postcodes or other means of geocoding the attribute data, or poor digitising with the geographic objects, this will need to be rectified. There can be a problem with using legacy data, which may contain errors, and gaps in the data. Tools and services are available to address the cleaning and geocoding of data.

6 The Need for Metadata

What is metadata?

"Data about data and usage aspects of it. This information will often include some of the following:

- What it is about
- Where it is to be found
- Who one needs to ask to get it
- How much it costs
- Who can access it
- In what format it is available
- What is the quality of the data for a specified purpose
- What spatial location does it cover and over what time period

When and where the data were collected and by whom and what purposes the data have been used for, by whom and what related data sets are available, etc.", (AGI Online Dictionary Definition - www.agi.org.uk/public/gis-resources/index.htm).

A metadata record should enable "the user of a dataset to understand the content they are reviewing, its potential and its limitations." (IGGI Guide to Principles of Good Metadata Management - Intra-Governmental Group on Geographic Information, 2002). People searching for datasets will be asking 'What GI is out there?' 'Where is it?' and 'How can I acquire it?' Metadata should provide the answer to these and help understand how best to use the data.

Why is metadata important?

The PAT 18 report (Social Exclusion Unit, 2000) drew attention to the fact that many policy makers are often unaware of what data is already available (and from whom), which could result in dataset duplication and wasted resources.

Metadata is important as it points to the existence of GI, and contains information regarding how it was created, is maintained, and enables the user to decide whether it is fit-for-purpose. Metadata is also important for enabling information sharing, which is crucial to a modernising government agenda advocating joined-up services and e-delivery.

Besides enabling suitable datasets to be found, metadata is necessary for data management. It should inform the data manager of when and how a dataset requires maintenance. It will also highlight potential information gaps, enabling the collection of missing datasets, and enhancements to existing ones.

Existing & planned metadata standards

Traditional metadata standards based upon the Dublin Core Initiative (http://dublincore.org/) do not include fields relating to the geographical component inherent in GI. The Dublin Core Initiative focuses upon the discovery aspect of metadata, i.e. it enables data to be searched for easily. It is the least sophisticated form of metadata, and as such is suitable for a wide range of purposes and business models. The e-government metadata standard is based around Dublin Core, but is not totally useful for GI. However the elements within it can be related to the UK geospatial data standard. For example, the 'coverage' element of e-gms (for limiting searches to information about a particular time or place) can be matched to NGDF elements about spatial location and temporal aspects.

The National Geospatial Data Framework (NGDF - www.ngdf.org.uk) is the UK geospatial data infrastructure, aiming to facilitate the availability of GI by enabling better awareness of data availability, improving access to data and integrating data by encouraging the use of standards. The Federal Geographic Data Committee (FGDC - www.fgdc.gov) maintains the equivalent US metadata standard. This has been adapted for use in other standards (e.g. Center for International Earth Science Information Network – CIESIN, and Australia New Zealand Land Information Council – ANZLIC). However it is very comprehensive with over 300 elements (or details to be given) about the data.

The International Organisation for Standardisation Technical Committee 211 (ISO/TC211 - www.isotc211.org) has prepared a draft international metadata standard (FGDC and NGDF have been actively involved in this). This new ISO standard is intended to be an improvement on existing standards, as it has optional metadata elements allowing for a more extensive standard description of GI to fit specialised needs.

How do you implement metadata creation?

The IGGI guide to Good Metadata Management (Intra-Governmental Group on Geographic Information, 2002) recommends creating roles to establish a metadata policy within an organisation. These include a policy champion at senior level (the policy owner that has authority and resources to implement the metadata system), a metadata steward (manages the metadata resources), and a compiler / maintainer (responsible for metadata record compilation, and maintenance).

If the above is too complicated, or the organisation is not large with few key datasets to manage, then a simpler approach will suffice. Essentially whoever creates a metadata record concerning a dataset, should be the person(s) responsible for creating this, or who uses it the most and has the greatest understanding about this dataset. They must try and include as much information as possible.

Besides establishing a metadata policy, it is important to ensure that metadata records are adequately maintained (for example, contact details listed within the metadata can quickly become out of date), as with the actual metadata policy (so that it is always in accordance with recognised standards). Regular internal audits of the datasets maintained should be undertaken to compile accurate metadata.

A technical metadata specification can appear very daunting. An easier way to approach metadata creation is to think of questions that are likely to be asked of the dataset. The following questions may assist:

- What does the dataset describe? (Title, geographic area covered, time frame).
- Who produced the dataset? (Contact details, and deputy contact).
- Why was the dataset created? (Useful for determining how it should be used).
- How was the dataset created? (Include accuracy, completeness, consistency, methods used to create, and date of creation).
- What entity and attribute information is available? (Names and definitions of features, attributes and attribute values).
- How reliable is the data? (Consistency of data capture).
- Is access available to a copy of the dataset? (Contact details, cost, available formats, copyright / intellectual property rights).
- Is spatial reference information available? (Coordinate system used in data creation, scale captured at, map projections, bounding coordinates).

The Open GIS Consortium is working to develop software tools to help with the metadata creation. The ISO standard does include guidance on how the metadata could be structured as XML. This is important for aiding efficient searching through data clearing houses, as XML adds contextual information about a dataset or information on a web page. GIS manufacturers are beginning to include tools to assist with creating metadata and XML tags. For example, in ArcGIS 8.1 the ArcCatalog can automatically create metadata about GI stored as XML.

7 Summary

Since the PAT 18 report into better information was published over two years ago, there is still evidence that data is not been taken seriously enough. This paper hopes to remind GI users of the pitfalls if data is not treated with the respect it deserves, and of ways to address any problems so that good GI becomes available. In addition to correcting problems with data, organisations need to ensure employees are

sufficiently trained in collection, maintenance, and use of GI. Consideration should be given to creating metadata records, as this will improve data maintenance and the dissemination of GI to other users. Once these issues have been considered, users of GI should reap the benefits of having clean, accurate, relevant GI used in the appropriate manner. Hopefully, then the vision of joined-up government and improved business development should become a reality.

8 Bibliography and Further References

Chainey, S. – 'The importance of geography: better information for tackling social inclusion', AGI Conference Proceedings 2000.

Crime Reduction – www.crimereduction.gov.uk

Federal Geographic Data Committee - www.fgdc.gov

Intra-Governmental Group on Geographic Information – 'Principles and Practice of Sharing and Trading Government Information', 2001.

Intra-Governmental Group on Geographic Information – 'The Principles of Good Metadata Management', 2002.

International Standards Organisation Technical Committee 211 - www.isotc211.org

National Intelligence Model – www.lancashire.police.uk/nimhome.html

Performance and Innovation Unit – 'Privacy and data-sharing: The way forward for public services', 2002. – www.piu.gov.uk/2002/privacy/report/index.htm

Social Exclusion Unit – 'National Strategy for Neighbourhood Renewal - Report of Policy Action Team 18: Better Information', 2000.