

Geodata: use with confidence?

JEREMY GILES¹

¹ British Geological Survey. (e-mail: jrag@bgs.ac.uk)

Abstract: Decisions about the geology of the urban environment are based on knowledge. This in turn is derived from professional expertise drawing on a range of data sources. These data sources are either newly acquired, specifically to address the problem in hand, or are items of prior information that were acquired for other purposes and are being re-used. Geodata are re-used constantly, but how can the user have confidence in the previously acquired information? A range of techniques are being employed at the National Geoscience Data Centre (NGDC) (www.bgs.ac.uk/ngdc) to ensure that the wealth of information held can be re-used to support the creation of new knowledge and hence address a range of issues in the urban environment.

Résumé: Les décisions sur la géologie de l'environnement urbain sont basées sur une connaissance qui est dérivée, de sa tour, de l'expertise dessinant sur une gamme de sources de données. Ces sources de données sont ou récemment acquises pour adresser un problème courant spécifiquement ou bien elles font partie de l'information déjà stockée qui était acquise pour un autre but et qui est réutilisée autre part. Géodata (les données géologiques) sont constamment réutilisées mais la question est comment l'utilisateur peut-il avoir confiance en information précédemment acquis. Une gamme de techniques est utilisée par le Centre National de l'Information Géoscientifique (www.bgs.ac.uk/ngdc) pour assurer que l'abondance d'information stockée peut être employée pour soutenir la création des nouvelles connaissances et par conséquent pour adresser une gamme de problèmes dans l'environnement urbain.

Keywords: geodata, database systems, geographical information systems

INTRODUCTION

The acquisition of new data as part of an investigation is expensive and it can be cost-efficient to reduce outlay by seeking to re-use existing data. By using this approach the costs can be reduced in a number of ways.

- The prior information can be used to provide a preliminary assessment (desk study), which informs the design of the proposed investigation.
- Where the data provide appropriate coverage they can be used in estimating the likely costs of a comprehensive ground investigation.
- High quality existing data can be used to validate, and even augment, the results of a new investigation.

The primary limitation on the re-use of prior information is user-confidence. How do those who collect, store, process and distribute prior information build user-confidence? This problem is faced daily at the British National Geoscience Data Centre.

NATIONAL GEOSCIENCE DATA CENTRE

For those who need to find out more about the subsurface of any area in Great Britain, whether for academic or commercial purposes, the most accessible and comprehensive collection of information is available at the National Geoscience Data Centre. This national collection is part of the British Geological Survey (BGS) and its main purpose is to support the Survey's own mapping programme. The National Geoscience Data Centre resources comprise data gathered or generated by the British Geological Survey or its precursor organizations, as part of the national strategic mapping programme, together with data provided by numerous external organizations both in academia and across a wide range of industrial sectors. The National Geoscience Data Centre is one of a suite of data centres managed by the Natural Environment Research Council (NERC), covering all aspects of the environmental sciences, including the atmosphere, oceans, water resources, ecosystems, and the geosciences. Their collective purpose is to provide access to factual baseline data for consultation and re-use by those studying the evolving environment.

The collections of the National Geoscience Data Centre cover all aspects of the geosciences and include digital data, paper records and physical materials. The digital data include:

- digital database systems;
- geographic information systems;
- 3D models; and
- scanned digital image collections including borehole logs, geological sections, photographs and samples.

Many paper records are generated in-house by BGS scientists (Figure 1). Other documents may be supplied to the National Geoscience Data Centre under the terms of various items of legislation, such as the Water Industry Act 1991. However, the vast majority are donated voluntarily by their creators, who recognize the value and importance of these national collections. Data types include:

- Borehole records;
- Site investigation reports;
- Geophysical logs;
- Shaft logs;
- Section descriptions;
- Maps;
- Plans;
- Mine plans;
- BGS field slips;
- BGS field notebooks;
- Coal mining records;
- Onshore and offshore oil industry records;
- Images.



Figure 1. A selection of NGDC documents

The physical materials collections (Figure 2), like the paper records, are largely donated on a voluntary basis. However, regulations developed from legislation such as the Petroleum Act 1998, provide for information to be added to the National Geoscience Data Centre as a statutory requirement. Data types include:

- Borehole core;
- Rock and mineral specimens;
- Thin sections of rocks and minerals;
- Rock and sediment core sample photographs;
- Biostratigraphical (fossil) material.



Figure 2. NGDC Core Store

The National Geoscience Data Centre holds a wealth of nationally important datasets. In managing these it aims to:

- Manage all data/information in accordance with NERC and BGS Data Policies;
- Manage all information in appropriate environments to safeguard and enhance its long-term potential;
- Ensure that appropriate metadata are provided for datasets and records;
- Provide digital indexes to actively-used datasets;

- Create a system in which all records, in whatever format or media, can be found rapidly;
- Build users' confidence by creating and maintaining validated and verified datasets to agreed standards; and
- Provide tools that enable geoscientists, both inside and outside BGS, to use BGS information with confidence.

The majority of the data held by the National Geoscience Data Centre are freely available for consultation and re-use by the general public, academia or industry. Some data items or datasets are held "in-confidence" for a specified period prior to their release.

RE-USING GEODATA

Data collected and interpreted for one purpose can commonly be re-used for another purpose, provided that a number of criteria are fulfilled. The re-user needs to be confident that:

- The original purpose for which the data were collected is clearly understood;
- The data were originally acquired in a professional manner, using what was considered to be good practice at the time of collection;
- The data represent an accurate record of the results of the investigation of the original project area;
- The data have been managed in a professional way, according to good practice;
- The data are easy to find, acquire and comprehend, and are available in a suitable format.

Original Purpose

It is important to know and understand the original purpose for which an item of information was collected. Such background information should tell the potential re-user of a dataset not only what was collected but also how it was collected. Boreholes drilled for hydrocarbons exploration require different techniques and follow different standards and best practice to those drilled as part of a geotechnical study. The datasets containing the results of the drilling are very different and are presented in different ways. Both include a wide range of valuable data that can be re-used for a range of purposes, but knowledge of the original purpose and specification of the drilling project is an important attribute of the borehole record.

Professionally Acquired

To help ensure development of an accurate understanding of the subsurface a well-designed investigation, using best practice, should be undertaken during the data acquisition phase. As with any form of human endeavour the quality of the outcome is related to the methods applied during the investigation and the skills, knowledge and professionalism of those undertaking the work. The methods used are normally documented in the report of the investigation, but they might only reference the appropriate code of practice or standard applicable at the time of the investigation. For example, the code of practice for site investigations in Great Britain (GB) has evolved through a number of distinct versions. The first incarnation was CP 2001 published in 1957. This was followed by the publication of the British Standard 5930 in 1981, which was reprinted several times, with a significant revision in 1999. When reviewing existing site investigation information it is essential that the data user should know which version of the code of practice was in use at the time of the investigation. The options are summarized in Table 1. This, of course, leaves some scope for ambiguity – during the transitions between codes of practice, was a particular firm still using the old standard or had they adopted the new standard at a given date?

Table 1. Summary of code of practice for site investigations in GB

Version of Code of Practice	Date Range
No published code of practice	Prior to 1957
CP 2001 : 1957	1957 – 1980
BS 5930 : 1981	1981 – 1998
BS 5930 : 1999	1999 – to present

Lack of confusion is only guaranteed if the data to be re-used are still attached to the original report, or reference its location. Such relationships enable the context of the data to be checked if there is ambiguity. Incomplete transcription of records into the digital environment can make the digital version more difficult to re-use than the paper original, where it is still a component of the report.

Identifying the methods adopted during the investigation is comparatively straightforward compared to the difficulties of assessing if those methods were applied professionally by the team producing the report. Experienced users acquire an instinct regarding whether or not an organization consistently produces reliable information. However, this is virtually impossible to quantify and, even more so, to document. Invariably this means that re-users must be circumspect when dealing with data gathered by organizations with which they are unfamiliar.

Accurate Investigation

Assessing whether a previous study has produced reliable results can be more straightforward. The basic approach is to consider whether the data being re-used are consistent with data collected independently from other sources. In urban areas there may be a considerable density of data available for re-use. For example, the National Geoscience

Data Centre contains borehole records for the centre of the city of Glasgow in the UK with a density in excess of 200 boreholes per square kilometre. Other city centres have similar densities of borehole coverage, the records having been acquired from various independent sources. Considering these in conjunction with each other it is possible to assess whether individual records are consistent with their near neighbours. This process can provide a degree of confidence in the available data, if they are consistent. This mutual validation technique can be very powerful, but of course its value is limited to areas with high densities of information. Outside the centres of major cities it is more difficult to assess the quality of isolated boreholes or groups of boreholes. Additionally, potential re-users must remain aware that, because of the nature of geology, adjacent boreholes, even just a few metres apart, may reveal entirely different sequences.

Where syntheses of geological information exist it is possible to test whether the borehole records are reasonably consistent with them. Geological maps synthesize the available information from disparate sources to provide the best geological interpretation of an area at the time of publication. The reliability of the map interpretation is, of course, also dependant upon the density and quality of available data and therefore will vary in accuracy from place to place. However, geological maps will normally be constructed on the basis of a higher than average information density in urban areas – the very areas where development is most common.

With the development of three-dimensional geological modelling it is possible to generate ‘virtual boreholes’. The British Geological Survey, in common with many other geological survey organizations, is starting to produce and market three-dimensional geological models. Selected urban areas have such geological models under development. These offer great potential, including the ability to create a ‘virtual borehole’ record at any point within the model. These virtual records can be used to test data before re-use. Actual borehole records can be tested against the model to see if they are consistent with the predicted (virtual) borehole at any specified location. If there is significant inconsistency between the borehole data and the model, both might need reviewing.

Well Managed

Once geodata have been created and used for their initial purpose they need to be transferred to a suitable data management system if they are to be of potential future use. Leaving potentially valuable datasets to look after themselves is not a reliable option. Storage of all forms of data incurs a cost, even when data merely sit in a filing cabinet. Typical site investigation data will cost about £4 (6) a year per report to store in quality office space and about £2 (3) a year per report in commercial storage. Leaving such data in a garage or cellar may appear cheaper, but data stored in this way commonly suffer damage. Figure 3 shows a vermin-damaged site investigation report that was recovered by National Geoscience Data Centre staff from an out-building adjacent to the offices of a major site investigation company.



Figure 3. Vermin-damaged report

Data Management and the related disciplines of Records and Archive Management have long established traditions of best practice, which should be followed. One of the best summaries of the role of a geological data manager is by Lowe (1995). The International Organization for Standardization (ISO) has now published a standard for records management (ISO 15489). In the UK the National Archive has been formed by merger of the Public Record Office and the Historical Manuscripts Commission, and it publishes extensive best practice documentation on its website (www.nationalarchives.gov.uk). Other government bodies, such as the Intra-governmental Group on Geographic Information (IGGI), also produce related and useful best practice documentation. The IGGI website contains a number of relevant best practice guidelines including *The Principles of Good Data Management* and *The Principles of Good Metadata Management* (www.iggi.gov.uk).

Feineman (1992) identified what he called the ‘*Dimensions of data management*’. These are a series of properties of a dataset that describe it and enable a potential user of the data to determine whether they are fit for the intended purpose and help to build user confidence. The eight dimensions are:

- Completeness – *All potentially available data are actually available;*
- Accuracy – *The dataset is error-free, or the error limits of the data are known and documented;*
- Fidelity – *The computerized representation of information reflects our actual understanding of the original raw data;*
- Quality – *The data are preferred because of exceptional completeness, accuracy and fidelity;*
- Lineage – *The original source of data is known, as well as details of all subsequent processes and transformations;*
- Timeliness – *The data represent the current state of knowledge, or the state of knowledge at the time of data collection/synthesis is recorded and described;*
- Accessibility – *The data store can be located, and has facilities to allow data movement to and from the repository as needed;*
- Security – *The data, and their related documentation are protected from unauthorized access, inappropriate use and partial or total loss.*

These dimensions are now largely subsumed into a comprehensive metadata description that would be compliant with ISO 19115 – metadata for describing geographic information.

Data Management policies and procedures ensure that data on all media are treated as a valued resource (IGGI 2005). Implementing good data management practice will potentially give many benefits to different communities within the data management life cycle.

The benefits to Data Suppliers are:

- An increased confidence and trust that their data will be used according to their agreed conditions of use;
- Provision of a clear understanding of the onward use of their data, documented formally in a Memorandum of Understanding signed by both supplier and user.

The benefits to Data Brokers/Intermediaries are:

- Better quality, harmonized and coherent data resulting from the use of common definitions, including geographical references, formats, validation processes and standard procedures;
- Better care of the data holdings through the use of effective data policies and best practice guidance;
- Better control over the data empowered by the clear definition and use of the procedures for the care of data.
- Improved knowledge and understanding of data holdings, their availability, interpretation and use, with subsequent reduction of the risk of misuse, duplication or loss, through better cataloguing, metadata and, in time, better access to data via an integrated data environment;
- Improved business processes, including better and more efficient use and re-use of data, and the standardization of datasets;
- Increased confidence that the organization complies with statutory and non-statutory obligations, by the regular use of centrally coordinated, frequently updated guidance, codes of practice and training on legal, contractual and other obligations;
- More sensible and consistent data charges and conditions of use, resulting from clear pricing and dissemination policies that recognize the need for free access by appropriate customers whilst recovering the appropriate income from customers who seek to make commercial gain;
- An increasing confidence by the customer in the quality of the data managed and in the reliability of outputs that are produced.

The benefits to users and customers are:

- Improved awareness and understanding of what data are available for current and future use, resulting from better cataloguing and data archiving;
- Improved access to data, free from unnecessary obstacles, safeguarded from disclosure of personal information or infringement of legal and contractual obligations;
- Better quality and more timely information;
- Better value for money, resulting from clear, fair and consistent data charges and conditions of use, which recognize the need for free access by appropriate customers;
- Better exploitation of data generally, enabled by easier data exchange and integration with other harmonized data.

The principal goals and activities of data management are:

- Avoid re-collecting the same or similar data;
- Manage the lifecycle of data from creation to disposal;
- Develop, maintain and implement a data policy;
- Manage IPR;

- Maintain comprehensive metadata; and
- Manage the data quality.

Easy to Find

The cost advantage of re-using data decreases rapidly if the records take a long time to find or reformat so that information can be extracted from them. Looking for data can be expensive, slow and frustrating. A study in the oil and gas exploration sector by Peebler (1996) produced the following observation:

“Lack of basic data integration costs the average E&P professional a considerable amount of time. According to various estimates geoscientists and engineers spend from 20% to 30% of their total project time searching for, loading and formatting data.”

Similar figures can apply to anyone attempting to re-use information if the underlying data are not made accessible and easily available.

Discovery Metadata provide the primary tool for finding information collections easily. A number of standards exist for describing information collections in this way, but all of these should eventually converge towards ISO 19115. A typically example of an existing standard is the UK GEMINI Discovery Metadata Standard (www.govtalk.gov.uk/schemasstandards/metadata.asp), which is a defined element-set for describing geo-spatial, discovery level metadata within the United Kingdom. The standard is the result of a year-long collaboration between the Association for Geographic Information (AGI) and the e-Government Unit, with additional representation from national and local government, and the academic community. Version One of this standard was published in October 2004.

First compiling and then managing discovery metadata are challenging tasks, but they are essential activities for any information repository. Good metadata management provides a wide range of potential benefits (IGGI 2004).

The benefits to Information Managers are:

- Provide an asset register of the key datasets managed;
- Provide a mechanism for maintaining and auditing the key datasets managed by the organization;
- Facilitate communication of the extent and range of the datasets;
- Prevent the re-creation of a dataset previously compiled by another organization;
- Facilitate the enhancement of an existing dataset;
- Facilitate compliance with data- and information-related legislation, regulations and directives;
- Enable details of the context in which data were collected to be retained with the dataset, so that appropriate and meaningful re-use of the data can be made in the future;
- Reduce the risk of datasets being devalued by knowledge about them being lost when key staff members move on.

The benefits to data brokers/intermediaries are:

- Enables a clearer understanding of the opportunities for developing value-added products;
- Facilitates the integration of diverse datasets to produce new value-added products;
- Empowers businesses and citizens to exploit datasets for information and knowledge that can support economic growth.

Benefits to users and customers are:

- Enables provision of gateways to search metadata by theme;
- Faster and easier discovery and acquisition of information.

The principal activities of good metadata management are:

- Establish a metadata policy;
- Adopt a metadata standard;
- Compile metadata;
- Publish metadata; and
- Maintain the currency/accuracy of the metadata.

The National Geoscience Data Centre has compiled comprehensive discovery metadata entries, which are actively maintained and currently published (October 2005) on the following sites:

- The British Geological Survey website – www.bgs.ac.uk/discoverymetadata/home.html
- The NERC MetaData Gateway – www.nmp.rl.ac.uk/
- gigateway – www.gigateway.org.uk/default.html

Once a collection of geoscience data has been located and its potential value estimated using discovery metadata, it should be possible to explore the information spatially. This requires a range of skills and resources, as the individual data items within the dataset need to be indexed digitally, including a standard spatial reference, and then published in either a desktop or web-based GIS. Several examples of such systems exist. The National Geoscience Data Centre publishes its key datasets on the GeoIndex (www.bgs.ac.uk/geoindex/home.html). This system permits the user to see index-level metadata about individual records within numerous datasets. For example the index of borehole records provides access to the following type of information (using an example from the Nottingham University Campus):

- Unique Borehole Identifier – SK53NW65
- Name – Nottingham University Library Area Borehole Number 7
- Grid Reference – SK5423738252
- Precision – Known to the nearest 10 metres
- Length – 10.21m
- Earliest data known to the NGDC – 1967
- Site investigation report cross reference – 11683
- Paper record held at – NGDC Keyworth

Such detailed spatial index metadata allow potential users to make a rapid assessment of whether the data are of potential value and could potentially be re-used.

Once such an assessment has been made and it has been decided that the data are of potential value it needs to be straightforward to obtain the information. The BGS operates a service called GeoRecords, which complements the GeoIndex. This service provides access to:

- Borehole records
- Coal Authority records
- Site investigation records
- Geophysical data
- Waste site reports
- Geological maps
- Geologists' field slips
- Geologists' field notebooks.

A further service provided through the National Geoscience Data Centre caters for the online ordering of boreholes (www.bgs.ac.uk/boreholes).

CONCLUSION

Geoscience data are created in great volumes by a range of industries, which spend large sums of money every year. Such costs can be reduced by exploiting existing information. However, this requires the infrastructure to be in place to acquire, manage and distribute such information in a way that gives potential users confidence. In Great Britain the British Geological Survey has established the National Geoscience Data Centre to meet both its own needs to manage and re-use existing information and to provide a range of facilities that benefit British industry.

Acknowledgements: I thank my colleagues in the National Geoscience Data Centre for their help in the preparation of this document, in particular Dr D. J. Lowe, for improving the quality of the early draft, and my multilingual secretary Ms J. Pakkanen for providing the French résumé. The paper is published with the permission of the Executive Director of the British Geological Survey (NERC).

Corresponding author: Mr Jeremy Giles, British Geological Survey, Keyworth, Nottingham, NG12 5GG, United Kingdom. Tel: +44 1159363220. Email: jrag@bgs.ac.uk.

REFERENCES

- BRITISH STANDARDS INSTITUTION 1981. Code of practice for site investigations: BS 5930. London, BSI
- BRITISH STANDARDS INSTITUTION 1999. Code of practice for site investigations: BS 5930. London, BSI
- FEINEMAN, D.R. 1992. Data Management: Yesterday, Today and Tomorrow. Presentation to the PETEX '92 Conference, 25 November 1992, London.
- INTERNATIONAL STANDARDS ORGANISATION 2001. Information and Documentation – Records Management – Part 1: General: ISO 15489-1:2001, ISO
- INTERNATIONAL STANDARDS ORGANISATION 2003. Geographic Information - Metadata: ISO 19115:2003, ISO
- INTRA-GOVERNMENTAL GROUP ON GEOGRAPHIC INFORMATION 2004. The principles of good metadata data management 2nd Edition. London, ODPM Publication.

- INTRA-GOVERNMENTAL GROUP ON GEOGRAPHIC INFORMATION 2005. The principles of good data management 2nd Edition. London, ODPM Publication.
- LOWE, D.J. 1995. The geological data manager: an expanding role to fill a rapidly growing need. In: Giles, J. R. A. (ed) 1995, Geological Data Management, Geological Society Special Publication No 97, 81-90.
- PEEBLER, R. 1996. Extended integration the key to future productivity leap. Oil and Gas Journal May 20, 1996; Vol. 94; No. 21