Database design in geochemistry: BGS experience

J. S. COATS & J. R. HARRIS[1]

Minerals Group, British Geological Survey, Keyworth, Nottingham NG12 5GG, UK
[1] *Present address: Nottingham City Hospital, Hucknall Road, Nottingham, UK*

Abstract: The Minerals Group of the British Geological Survey (BGS) has been successfully using a relational database for over seven years to hold geochemical and related locational and geological data collected by the DTI-sponsored Mineral Reconnaissance Programme (MRP) since 1972. Operational experience has shown that several problems and deficiencies of the database are the result of the lack of normalization of the original data model. A new data model, derived from analysis of the data collected by the MRP and other geochemical programmes, is presented and solves many of these problems. It is designed to hold all BGS geochemical data for the UK landmass and is currently being implemented. It is presented as a contribution to a wider debate on generalized data models for geochemistry.

All databases attempt to model reality and hold information about their world of interest. These models can vary in complexity and, also, change and evolve with time. This paper reflects the experience of the Minerals Group of British Geological Survey (BGS) in using a simple database design for the past six years and the design principles employed in creating a new conceptual schema or logical model for a geochemistry database.

The Minerals Group of BGS, or its predecessors, has been carrying out geochemical exploration surveys for the past 30 years, initially in connection with uranium exploration in the UK but, for the past 20 years, conducted under the Mineral Reconnaissance Programme (MRP). The MRP is funded by the Department of Trade and Industry with the aim of encouraging exploration in the UK by mining companies. BGS has completed 127 reconnaissance mineral exploration projects over this period and these have been published as a series of MRP reports (Colman 1990). Notable successes of the MRP have been the discovery of the Aberfeldy baryte deposits (Coats *et al.* 1980) and, more recently, gold in south Devon (Leake *et al.* 1992). The MRP has always been multidisciplinary in nature, with input from geologists and geophysicists, but this paper is chiefly concerned with geochemical and related geological data. Geophysical data collected as part of the MRP have been incorporated into an index-level geophysical database run by the Regional Geophysics Group of BGS. Geochemical work has also been carried out in several areas outside of those described in MRP reports. Unpublished internal reports may exist for the areas or, if the study was inconclusive, the data left in paper or computer files. One of the primary aims of the MRP database project has been to make these data accessible to BGS staff and to exploration companies.

The greater use of geochemistry in mineral exploration over the 30 year period has been made possible by the increasingly automated methods of analysis, which have changed from essentially manual methods, such as colorimetric, DC-arc emission spectrography and atomic absorption spectrometry, to the rapid, multi-element X-ray fluorescence and ICP spectrometry techniques in use today. Data recording methods have also changed over this period, with the early results being recorded on paper, then on pre-printed results sheets for data entry on punched cards and, finally, produced by computers incorporated into the analytical equipment and communicated over a site network or communication link.

The quantity of geochemical data collected by the MRP and related projects is estimated to be about 1.6 x 10^6 analyses by various methods on 125 000 samples. An accurate estimate is difficult to calculate because of the lack of a complete index to the number of analysed samples but records of the BGS analytical laboratories allow this approximate estimate to be made.

Real system

The real system modelled by the database is specific to geochemical mineral exploration but this area of interest is not very different from geochemical mapping or lithogeochemistry. These methods are described in textbooks, for example the volumes comprising the *Handbook*

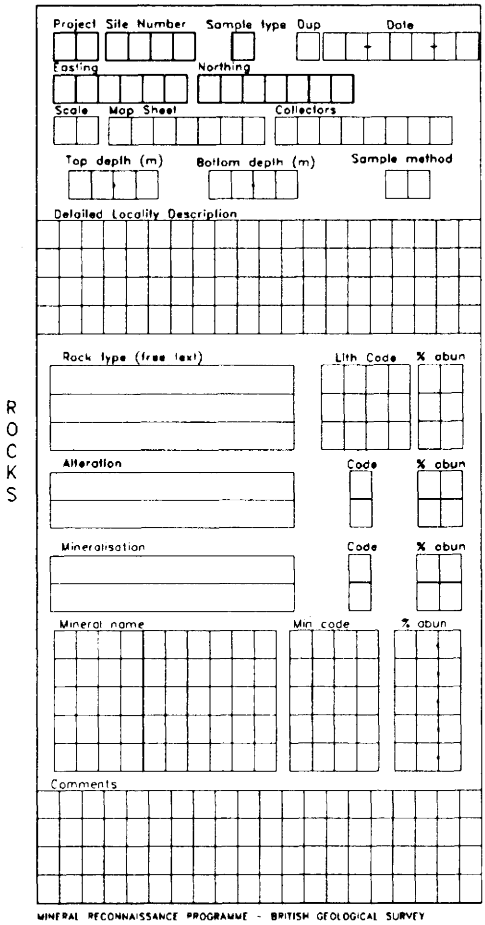
From Giles, J. R. A. (ed.) 1995, *Geological Data Management*,
Geological Society Special Publication No 97, pp. 25–32.

of Exploration Geochemistry (the latest in this series (Kauranne *et al.* 1992) describes regolith exploration geochemistry). Very briefly, the sampler moves to a sampling site and, using a variety of methods, such as digging with a spade, using a portable drill (for overburden samples) or collecting a rock sample with a hammer, collects a geochemical sample which is given a unique sample number. This sample is then sent to the laboratory for analysis. The results are reported to the collector who uses the information, together with details collected in the field, to predict the occurrence of mineral deposits.

The most important item recorded in the field is the sample location but, typically, other relevant details of the site are recorded, particularly those that are likely to affect the geochemistry, such as depth for an overburden sample. Samples may differ in colour and texture and in some cases this can be important information used in the interpretation. A leached, white sample will be significantly different in chemical composition to the adjacent, red, iron-enriched horizon. Information is therefore collected about the sample as well as the location. BGS has been collecting this field information since 1970 on field cards that were designed to record the relevant sample information and location on a simple pre-printed form held in a small ring binder or Filofax (Fig. 1 shows the latest version for rock samples). Three forms were used in the original system, one each for soil, rock and drainage samples. These forms replaced the traditional field notebook, and information on the forms was intended to be stored or used in a computer.

Samples are normally sent to the laboratory in batches, which are prepared in the same way and analysed by the same methods. Nearly all geochemical laboratories record a batch number that is used to identify the group of samples through the analytical system and this batch number is a useful index to the samples. Chemical elements can be determined by a variety of methods and the results can be different depending on the method. These methods may also have different detection limits, bias or interferences. It is therefore important to record the method of analysis of each batch. This extremely brief introduction gives some indication of the data that need to be recorded in a geochemical database.

The flow of data in the MRP system can be shown in a simplified flow diagram that gives an overview, from sample collection in the field to final MRP report (Fig. 2).



ROCKS

Project Site Number Sample type Dup Date

Easting Northing

Scale Map Sheet Collectors

Top depth (m) Bottom depth (m) Sample method

Detailed Locality Description

Rock type (free text) Lith Code % abun

Alteration Code % abun

Mineralisation Code % abun

Mineral name Min code % abun

Comments

MINERAL RECONNAISSANCE PROGRAMME - BRITISH GEOLOGICAL SURVEY

Fig. 1. Field data collection form for rock samples (1991 version).

MRP database

The current MRP database was designed in 1986 by K. A. Holmes and J. S. Coats, and subsequently enhanced in 1989 by J. R. Harris with the addition of a user-friendly front end. The introduction of a relational database management system (ORACLE) in 1986 allowed data files, which has been previously kept in a flat-file database in an internally designed programming system G-EXEC (Gill & Jeffery 1973), to be managed in a modern software environment. The MRP model was based on a one-to-one correspondence between sample type and database table. Thus, there are tables for the four main types of samples collected: rocks, drill cores, soils and drainage samples (Fig.3). The last table contains details about several physical

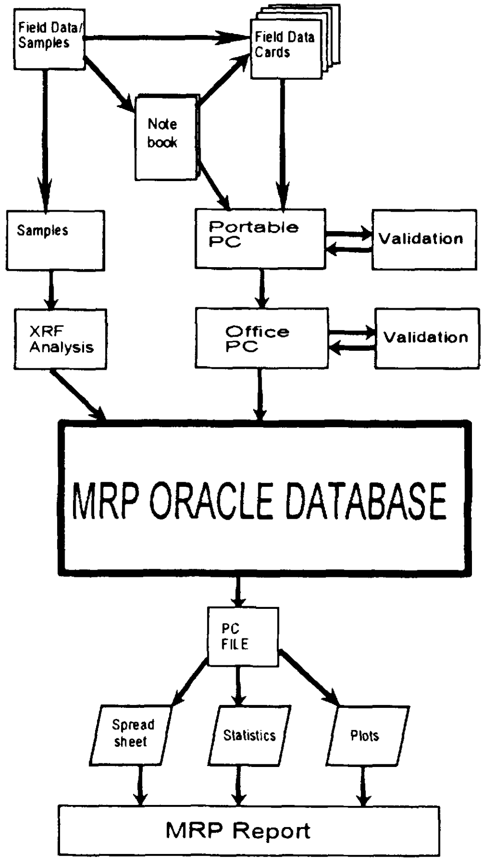


Fig. 2. Flow diagram for data and samples collected by the MRP.

samples (steam sediments, pan concentrates and water samples) collected at the same site. Two other tables hold the borehole locations (BHLOCAT) and the analytical results (ALLEL).

The borehole location table (BHLOCAT) contains the locations of the boreholes and such information as the elevation of the collar and the date of drilling. The relationship to the bore sample table is maintained by the borehole's name. A better primary key would be the BGS registered boreholes number but at the time of database design this was not available in a database table.

The ALLEL table holds the analytical results in a de-normalized format with a single column for each chemical element. To allow for determinations of the same element by different methods, the primary key includes the method of analysis. The relationship between the sample tables and the analytical table is one-to-many.

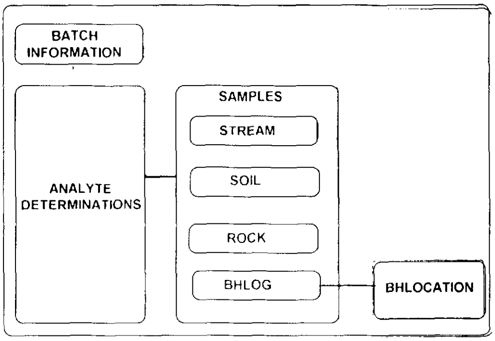


Fig. 3. Entity-relationship diagram for the MRP database model.

Each sample can be analysed by many methods and a single field sample can have a duplicate analysis by the addition of a duplicate key. The relationship to the sample table is maintained via a foreign key which is optional to allow for the entry of analytical data without the corresponding field data and, a field sample does not have to be analysed.

Access to the tables is via user views which enforce security. Screen forms have been designed to allow user data entry and updating. These forms also maintain a full audit trail of all updates and deletions of data. Bulk loading of data is performed by FORTRAN programs which ask the user to describe the format of the input data file and name the columns of the table to be loaded. A menu-driven front end written in VAX DCL allows the non-SQL (Standard Query Language) user to navigate the forms and data entry programs. As the number of tables is relatively small, most SQL retrievals are simple to construct and, with training, most users of the database can construct their own queries. Example SQL retrievals are given in a manual issued to every user.

Several deficiencies have been identified in the present model and many of these are the result of the lack of full normalization of the data tables. For example, in the STREAM table there is a text field that lists the minerals identified in the heavy mineral concentrate. Because of the lack of normalization, this field can only be searched inefficiently as a variable length text field. Also, problems are encountered with trying to perform searches for 'GOLD' and, also, for synonyms such as 'Au' or 'AU'. This repeating group error is common in the field cards, where a number of items need to be recorded about the sample or site. Another example is catchment geology, which can be composed of several rock types, but is only recorded here as major and minor rock type. Another deficiency concerns the inability to locate and validate all the different types of samples within a certain area. The sample locations are split between four sample tables, SOIL, ROCK, STREAM and BHLOCAT, and searching can be laborious.

A particular difficulty is found in the ever-increasing width of the tables. In the existing database if a new element is determined on a batch of samples, for example a few samples are analysed for iridium, then a new column has to be added to the ALLEL table. In addition, as part of the design of the new model it was decided that each element determination also required a qualifier as many laboratories report determinations as greater than or less than a specific value. This would cause the width of the ALLEL table to double and therefore exceed the limit of 255 columns per table imposed by ORACLE. A similar problem exists with the STREAM table where the width increased with every new attribute that was added. This causes problems when sizing database tables as ORACLE does not guarantee that nulls will not be stored if a row is updated.

Another problem with the existing database is that the coded data on the field cards have evolved since the time of their introduction in 1970. As these codes were not identical it is impossible to carry out a retrieval without specifying the year of collection and also knowing the coding system employed in that year. Retrievals on data going back over a period of a few years can be very difficult. It was therefore decided to translate all past data into a new set of comprehensive field codes, and these '1990' codes would be adopted as the domains for the data analysis.

New model

The aim of the new model was to describe accurately all BGS geochemical data for the UK landmass (excepting Northern Ireland) in order to facilitate the management of the data in a relational database. The new model was designed by the authors (Harris & Coats 1992) to meet this requirement and to overcome the deficiencies identified in the MRP system. During the redesign it was noted that the fundamental requirements and underlying structure of several other BGS geochemical datasets were very similar and they could be combined in one integrated BGS Geochemistry database. It was therefore decided to enlarge the aim of the new database from just MRP data to all BGS UK landmass geochemical data.

A fundamental, but unstated, requirement of the first database design was that it should hold as many of the geochemical data as possible. The data model produced by the redesign is based on a superset of the data that exist and is, therefore, data-driven. A complete set of the field cards and the associated field handbooks, dating back to 1970, was collected. In all, ten sets of field cards were identified and with the extensive redundancy and complexity of the complete dataset it was decided to translate them all into a new set of field codes. Thus all the data would be stored in a single format that would make retrieval of a whole dataset more efficient and straightforward. A dictionary was compiled in the form of an ORACLE table containing every possible value of every field on the different cards and the values translated into the new set of '1990' codes (Harris *et al.* 1993). These new codes were adopted as the domains for the data analysis that followed and the attributes for each domain derived from the codes. These attributes were grouped into the initial relations similar to those in the original MRP database and then normalized to third normal form or higher.

A few of the general principles used in designing the new codes should be mentioned before describing the attributes. Because of the difficulties of validating free text and to enforce the recording of data in a format compatible across many sampling teams, coding of data was used wherever possible. The codes employed were those used in the existing forms or, where these could be shown to be defective, new ones were established. Hierarchical coding schemes were used wherever possible and preferably those published by other organizations or experts.

The entity–relationship (E–R) diagram for the logical model is presented in Fig. 5. The diagram

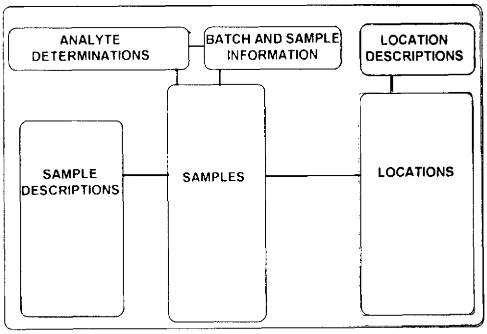


Fig. 4. Entity–relationship diagram for the geochemistry data model.

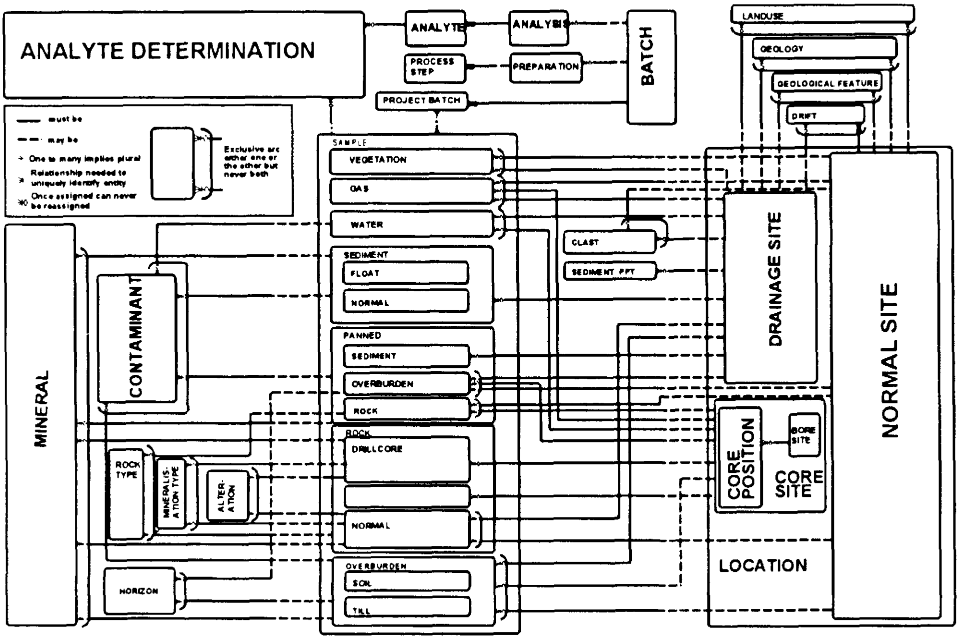


Fig. 5. Logical areas of the geochemistry data model.

can be divided into six logical area: locations, location descriptions, samples, sample descriptions, batch information and analyte determinations (Fig. 4). Each one of these areas will be considered in turn, giving the attributes recorded for each entity and the relationships. Detailed descriptions of the entities and domains are given in Harris & Coats (1992).

Locations

Location is one of the key entities of a geochemical sample. Three subtypes of location are recognized: normal, drainage and core sites. An attempt was made to combine these into one entity but the information recorded at a drainage site, such as flow conditions or stream width, is very different to that at a rock sampling location. Borehole core locations are also different in that they are located by reference to the distance down the borehole, as well as the drill collar location. Normal and drainage location entities have *project* and *site number* as the primary key and the following attributes: National Grid easting and northing, grid accuracy, grid derivation, local grid east and north, elevation, top and bottom depth, map scale, map sheet, gamma count, solid angle and gamma environment, detailed locality, comments, collector and date collected. Additional attributes recorded for normal sites are relief and profile drainage. Drainage sites have the extra attributes of drainage type, flow conditions, stream order, catchment area and weather.

Borehole sites are a special type of location that have National grid coordinates, a borehole name and number, drilled length, inclination, azimuth, drilling organization and name of core logger. The BGS standard reference number for the borehole is also added so that the table can be linked to the main BGS borehole database. The table linked to this location entity is core position, which has the project code and site number for the sampling site within the drilled length. This forms an alternate key to the table along with the primary key of borehole name, borehole number, top and bottom depth.

Location descriptions

Linked to the normal and drainage site locations are tables describing the surrounding area for land use, geology, geological features (such as faults or mineral veins) and drift. Land use uses a domain of 102 hierarchical codes derived from the 1966 Land Utilisation Survey and includes both natural and industrial uses. The latter are clearly important in urban areas and in identifying possible contamination. The geology table uses existing BGS hierarchical codes for lithology (Harrison & Sabine 1970) and the attributes, stratigraphic age and relative abundance. The drift table uses a domain derived by classifying the commonly used units appearing on the margins of geological maps. It is expected that this area of the database will be revised when other subject areas of the BGS data architecture are defined in detail.

Linked to the drainage site are tables containing information on clast composition and abundance, and sediment precipitate, both of which are important in identifying possible sources of geochemical anomalies. Originally, the database model linked contamination to the drainage location entity, but it was felt that contamination should be linked to actual contamination of the sample, rather than to potential contamination at the site. Clast composition can also be linked to a normal site to allow for the recording of clasts observed in overburden profiles.

Samples

Several sample types can be collected from one location and the relation to the location entity must be enforced. In other words, every sample must have a location but not every location has to be sampled. Also at a drainage site, samples of stream water, stream sediment and heavy mineral concentrate can be collected from the same location. The subtypes of sample are identified by the sample type code and the standard identification has been inherited from the original 1970 definition employed on the early fields cards (S = soil, C = stream sediment, P = panned concentrate, T = till, D = drill, etc.). This has some deficiencies but the difficulty of recoding many thousands of samples made this impossible to change. The sample inherits the project code and site number from the location table and, when combined with the sample type and duplicate codes, forms the primary key to the table.

Sample descriptions

Entities linked to the sample are contaminant as described above, horizon, rock type (or lithology), mineralization, alteration and mineral. The BGS standard mineral codes are used for the mineral domain (Harrison & Sabine 1970) and these can be linked to many of the sample types. Using this technique the occurrence or content of all minerals in many different types of sample can be managed and validated in one table and not scattered as attributes through many tables. The entity 'Horizon' uses a domain derived from the Soil Survey (Hodgson 1976) to describe the horizon that is sampled and has the attributes of Munsell colour, soil texture and soil type.

Batch and sample information

Batches of samples are prepared and analysed and this subject area of the database forms a metadata index to the sample analyses. Samples are grouped into project batches which have the same project code and sample type. The maximum and minimum site numbers, and the total number of samples, are recorded. This allows summary information on the range of samples included in one batch to be retrieved. An entry in the batch table may be made up of many project batches, thus allowing samples from several project areas to be combined. The batch entity has a primary key, composed of the batch identity and the laboratory, and attributes of date registered, geographical area and locality, which is validated against the Ordnance Survey gazetteer. Using this field, searches such as 'the batch of rock samples collected in Aberfeldy in 1980' can be completed. The analysis table holds information on the date and method of analysis of the batch, and the elements or species determined by that method are held in the analyte table, which contains quality control data such as the limits of detection. Other quality control data, such as precision, can be linked to this table by the analyst.

Originally there was the intention to model the complete laboratory analysis and preparation system, but this is a complex logical area and does not offer many benefits in added information value to the aims of the database. A Laboratory Information Management System (LIMS), which interfaces the analytical instruments and communicates with the Geochemistry database, is more appropriate for this task and is currently under evaluation in BGS.

Analyte determinations

The analyte determination table is the largest in the database and it is estimated to contain six million element determinations when fully populated. Because each row contains only sample number, method of analysis, laboratory, batch identity, analyte and abundance, the table is fully normalized and contains no null abundance values.

Retrieving data in tabulated form from this kind of structure was examined in detail and can best be achieved using SQL by decoding on the method of analysis and the analyte and then grouping by the unique sample identifier and averaging the values to obtain one row per sample. The average has no effect as it is the average of a series of nulls and one value, and is used merely to obtain the desired format (Harris & Wild 1992). Determinations are all stored in the same units (ppm by weight) to maintain uniformity and to avoid confusion. Parts per million was chosen as the best compromise between percentages, which would involve many zeros, and parts per billion, which would involve very large numbers. The majority of elements have Clarke values in the 1–10 000 ppm range (Mason & Moore 1982). The external schema (the view of the database seen by the user) may have per cent oxide for the major elements but retain the concentrations in the logical schema in parts per million.

Referential integrity

Site and sample number are of crucial importance to the data model because they are inherited as the primary key of most of the tables. The first work reported after the data design (Harris & Coats 1992) was concerned with an analysis of the sample numbering systems that had been used by BGS (Harris *et al.* 1992*a*). These systems have shown remarkable consistency over the long life of the Survey and, apart from the temptation of every new recruit to BGS to design their own numbering system, have been very successful in mineralogy and petrology, geochemistry and geology. Separate, but very similar, systems have been established in each of these areas and, by using a single code to identify the numbering system, all the samples collected by BGS can be uniquely identified. Rules have been developed to police the numbering system and to renumber samples that break them. Because the database captures much of the analytical data before it reaches the geologist, it can enforce these rules and prevent duplicate sample numbers being entered. A full record can also be kept of any subsequent renumbering.

Batch information is also critical to the database design as it provides a management and quality assurance tool and, also, a metadata index to all the analytical data produced by BGS. Sample analyses without batch information may suffer from a variety of problems, such as missing analytical methods, unknown laboratories, variable precision or detection limits and even post-analysis ‘normalization’.

Discussion and conclusions

The logical model presented in this paper is capable of describing the geochemical data held by BGS for the UK land surface and thus meets its primary aim. With the addition of latitude and longitude, and further fields to describe seabed sediments, it should also be able to hold data pertaining to offshore data. Geochemical data from elsewhere in the world can be largely encompassed by the model (Jones & Coats 1993) but difficulties arise because of the lack of control over the site and sample numbering system. A relational database cannot allow duplicate primary keys in its tables and rules must be devised to add an extra attribute for the country of origin.

A difficulty with the new data model is that it does not include bibliographic references. This was not included in the logical data design and did not feature as a strong requirement of the users interviewed before proceeding to the physical design (Harris *et al.* 1992*b*). The numbering system allows the identification of the main project under which the sample was collected, but does not reference the report in which the result was reported. As a single reference may report many analyses and an analysis may be reported in several references, this is a many-to-many relation. The number of sample analyses to be held by the database is predicted to exceed six million, and the linking table to hold this relation would be very large. Other smaller tables such as the project batch are possible candidates but are logically difficult to join to references. As the requirement for a link to bibliographic references is only small and chiefly confined to lithogeochemical samples that form only a small minority in the database, the relation has not been implemented.

The geochemistry data model described in this paper is broadly compatible with the BGS data architecture (Logica 1991) but there are differences. In the geochemistry data model the site or location number is inherited by the sample (because geochemists usually collect at least one sample from every site they visit). In the BGS data architecture, site and sample numbers are different (because a geologist mapping an area visits many locations but collects only a few samples). This difference is not yet a problem because the databases are not closely linked but may be a difficulty in the future.

This paper is presented partly as a case history of seven years' experience in operating a relational database to hold geochemical data and, also, to present a new data model for geochemistry. Co-operation in other areas of geology, such as oil exploration, has led to the development of common standards for relational data models (Petrotechnical Open Software Corporation 1993) and this paper is a contribution to a future debate on standards for data models in geochemistry. At present, where large datasets from many different exploration companies have to be combined, only the simplest form of data structure, that of a flat-file data table of analytical results, is possible. The evolution of common standards for relational data models will allow much more sophisticated data integration and analysis.

The authors would like to thank their colleagues in BGS for their co-operation in this work and the Sandwich Students from the local Universities who have helped in the implementation of the database. The authors publish by permission of the Director of the British Geological Survey (Natural Environment Research Council).

References

COATS J. S., SMITH, C. G., FORTEY, N. J., GALLAGHER, M. J., MAY, F. & MCCOURT, W. J. 1980. Stratabound barium-zinc mineralisation in Dalradian schist near Aberfeldy, Scotland. *Transactions Institution of Mining and Metallurgy (Section B Applied Earth Science)*, **89**, B109–122.

COLMAN, T. B. 1990. *Exploration for Metalliferous and Related Minerals in Britain: A Guide*. British Geological Survey, Nottingham.

GILL, E. M. & JEFFERY, K. G. 1973. *A Generalised Fortran System for Data Handling*. Institute of Geological Sciences, Computer Unit, Bulletin **73/3**.

HARRIS, J. R. & COATS, J. S. 1992. *Geochemistry Database: Data Analysis and Proposed Design*. British Geological Survey Technical Report **WF/92/5** (BGS Mineral Reconnaissance Programme Report 125).

——, GLAVES, H. & COATS, J. S. 1992*a*. *Geochemistry Database Report 2: A Proposed Integrated Sample Numbering System*. British Geological Survey Technical Report **WP/92/5R**.

——, COATS, J. S., WILD, S. B., KHAN, A. H. & STEPHENSON, P. 1992*b*. *Geochemistry Database Report 4: The User Requirement*. British Geological Survey Technical Report **WP/92/12R**.

——, NICHOLSON, C. J. & COATES, J. S. 1993. *Geochemistry Database Report 6: Standardisation of Geochemical Field Cards 1970–1992*. British Geological Survey Technical Report **WP/93/20R**.

—— & WILD, S. B. 1992. *Using the Oracle V.6 Explain Plan Utility*. NERC Computing, No. 54, 22–26.

HARRISON, R. K. & SABINE, P. A. (eds) 1970. *A Petrological–Mineralogical Code for Computer Use*. Institute of Geological Sciences Report **70/6.**

HODGSON, J. M. (ed.) 1976. *Soil Survey Field Handbook*. Soil Survey Technical Monograph No. 5. Soil Survey, Harpenden, Herts.

JONES, R. C. & COATS, J. S. 1993. *Sources of Geochemical Data Collected by the International Division since 1965*. British Geological Survey Technical Report **WC/93/17**.

KAURANNE, L. K., SALMINEN, R. & ERIKSSON, K. (eds) 1992. *Regolith Exploration Geochemistry in Arctic and Temperate Terrains. Handbook of Exploration Geochemistry Volume 5*. Elsevier, Amsterdam.

LEAKE, R. C., CAMERON, D. G., BLAND, D. J., STYLES, M. T. & ROLLIN, K. E. 1992. *Exploration for Gold in the South Hams District of Devon*. British Geological Survey Technical Report **WF/92/2** (BGS Mineral Reconnaissance Programme Report 121).

LOGICA COMMUNICATIONS LTD. 1991. *British Geological Survey Data Architecture Study*. British Geological Survey Database Documentation Series Report.

MASON, B. & MOORE, C. B. 1982. *Principles of Geochemistry*. 4th Edition, Wiley, New York.

PETROCHEMICAL OPEN SOFTWARE CORPORATION 1993. *Epicentre Data Model Version 1.0*. Prentice-Hall, New Jersey.