TECHNICAL METADATA IN THE BRITISH GEOLOGICAL SURVEY

Peter G Robson
Database Administrator,
British Geological Survey

Summary. The British Geological Survey (BGS) has been collecting and storing data pertaining to the United Kingdom landmass and offshore areas for over 160 years. This has resulted in the accumulation of considerable quantities of data, notable more for its extensive diversity than outright volume.
Since 1986 the BGS has been using Oracle to assist in the management of the ever-expanding digital component of its data. With over 4400 distinct database tables, covering a wide and diverse range of geological topics, we needed a means of imposing some logical structure onto such a diverse collection of tables.
Our solution lay in the definition of 'databanks' - assemblages of all those database objects (but most particularly tables) which combine to represent one area of activity in the Survey. This definition has been tightly coupled with the Oracle data dictionary (dd), so that by using a combination of views into the dd, and triggers to enforce our model, we have essentially extended the Oracle model to improve the management of a diverse, heterogeneous collection of scientific and management data.

Introduction

The British Geological Survey is the UK national centre for earth science information and expertise, and exists as one of five major environmental surveys within the Natural Environment Research Council (NERC). The BGS is one of the world's leading geoscience organisations. It has a staff of almost 800, of whom 500 are scientists, many with international reputations in their fields.

The Survey was founded in 1835, the very first such national Survey. In many respects, other national surveys adopted the model of the BGS when they came to be established.

As currently constituted, the BGS follows two closely related activities – its 'core' tasks (broadly, the surveying and assimilation of the geology and geological data as it relates to the UK landmass and continental shelf), and secondly, contract work undertaken predominantly for government departments and agencies, both in this country and overseas.

This reflects the source of funding which the Survey receives – from government, through the NERC, with a shortfall which has to be made up by seeking contract work. The two activities are closely interdependent, as skills developed in the core programme are frequently used in subsequent contract work, and vice versa.

Given the time over which the BGS has been collecting data, together with the many very distinct activities which it has been involved in, the variety of data collected is vast, in both diversity and quality. Until relatively recently, most of this data was in paper form, but over the last twenty years an ever-increasing proportion of it has either been collected in digital form in the first place, or rapidly converted to digital once it has been collected.

As well as alpha-numeric data concerning geological information, the Survey also holds considerable quantities of sample materials (rock, soil, fossil, mineral, maps, images, vectorized geological lines, 3D models, library material, digital text etc). All of this material is meticulously catalogued, classified and stored. Because this index information is now being held digitally, the Survey is already realising the added value of efficient association between what were often disparate and disconnected activities within the organisation. This is one of the key benefits of using digital, and particularly relational dbms technologies.

The range of data collected and held by the Survey can be appreciated from the organisational structure of the BGS. The core activities includes classical geological land survey work, as well as hydrogeological studies, coastal and engineering geology, mineralogy and petrology, analytical and regional geochemistry, fluid processes and waste management, petroleum and marine geology, regional geophysics, global seismology and geomagnetism, and basin analysis and stratigraphy. Contract work may be carried out within any of these specialisms.

Organisational support activity falls into two principle areas, one being responsible for administration, personnel and finance, and the other forming an infrastructural supporting role, including the main IT services in all their various guises from hardware support, software, applications, database, GIS, and R&D. Also present in this group is the training section, publications and business development.

IT Developments in the BGS

As already intimated, the range of data held by the BGS is extensive, from the highly structured (e.g. tabular numeric data), as collected by the geophysicists, to the highly qualified and unstructured (e.g. text descriptions of specimens), as recorded by the field scientists, with every variation in type and quality between.

Consequentially, various digital methods were adopted to manage this data, where it existed in digital form. The highly numerate components of the Survey (geophysicists, hydrogeologists, geochemists) took early advantage of computing facilities, but the more conventional parts of the Survey started to use the technology much later. They came into a world dominated by Fortran running on PDP-11 machinery, in which the normal user interface was the teletype! Such devices were rapidly replaced by glass terminals accessing multi-user DEC machinery, but all data was formatted and held according to the familiar 80-character card standard. Retrieval of these data was invariably by Fortran programs, especially written for the purpose. There was no such thing as RAD or a 4GL in those (recent) days.

Geology is a science in which the spatial distribution of objects is of prime interest to the geoscientist. Until computer output could generate graphical representations of maps, sections, and 3-dimensional models, interest in computing technology within the Survey was somewhat muted. Such graphic output first appeared as printed output, then on crude glass screens. The ability of current software to present complex graphics interactively, integrating raster and vector, has stimulated enormous growth of computing technology in BGS, and consequently placed heavy demands on the storage and presentation of the large quantities of data required to service such processes.

The first major step forward in data storage and management came in the early 1980's, with the introduction of one of the first relational dbms products (Mimer). This made a very positive impression on the staff involved in data processing, indicating just how productive computing could be. Within a couple of years the Survey adopted Oracle as its general purpose dbms, following a standard set by the parent body NERC. The policy in those early days was to permit all users to have resource privilege. Of course, the inevitable consequence was an absolute proliferation of tables across numerous schemas. Many of these tables held key Survey data, but often as not they were inaccessible to all but those most closely involved in the use and management of that data. Culturally, staff were still not prepared to 'grant select on mytable to public' – this sharing of 'their' data was a cultural block still to be overcome.

There were three fundamental problems already emerging here – first, there was no way in which the general BGS scientific community could know what digital data existed in Oracle. Secondly, even if they knew, they would probably not have access to review that data. And finally, even if they obtained access to the relevant tables, there was often no way of knowing what the significance of the various tables and their attributes actually meant.

Despite these early problems, the benefits of dbms, and particularly rdbms, were absolutely undeniable. Nevertheless, it was becoming increasingly apparent that steps had to be taken to rationalise the use of Oracle. The group charged with introducing dbms into the Survey made certain recommendations, which were then validated by a short-term consultancy, brought in from the CCTA (Central Computer & Telecommunications Agency). Basically, the BGS required an overall Data Architecture as a means of taking a high-level overview of the entire data holdings of the Survey, and thus integrating the digital and non-digital as one seamless resource.

Data Architecture

In 1990 a contract was awarded to Logica to assist the Survey in developing a Data Architecture. This was a significant learning experience for both parties, but established emphatically that although geoscientific data may be very different from that data more normally processed in large commercial organisations (as typified by the usual Logica customer), this data was nevertheless subject to exactly the same rules of analysis and design that pertains across the entire data processing industry.

During the project, BGS adopted the use of CASE tool technology. The chosen product was Systems Engineer from LBMS. At that time we required a product which would run on a PC, and the Oracle offering was only able to run on a Unix workstation.

The benefits of this project are ongoing, and have continued to accrue over the subsequent years. They have emphasized and reinforced the crucial importance of data as a corporate resource, irrespective of which department or division in the Survey was ultimately responsible for it. Due to a significant training programme (which has been maintained), a number of key staff have become familiar with the principles and concepts of data modeling and analysis, which again feed back into the primary benefit. Even more importantly, senior management gave, and continue to give full support to the concept of corporate data, and to the elimination of isolated pockets, or islands of data, being held secretively by disparate groups.

Steps were taken to implement this view of data within Oracle. The majority of information recorded in Oracle was index data, pointing to the existence, location, and description of other material, whether as reports, specimens, maps, or other physical items. A policy was adopted whereby all data recognised as being of corporate value was to be held in the same Oracle schema, identified by the name ‘bgs’. To persuade people to have their data assimilated into the corporate schema, a new, high performance database server was installed specifically for the corporate data, thus providing an incentive for staff to migrate their data.

Following on from the Logica project, different groups in the Survey started to model their own data, using the corporate standard CASE tool. These individual models were then submitted to the BGS database administrator, who would integrate them into one single all-encompassing schema. Although this schema began to take on a very complex appearance, it was always possible to decompose it to its constituent sub-models, according to the geoscience discipline from which they were derived, but only within the confines of the CASE tool. As this was a single user environment, the benefits of that decomposition could not be seen by the end user community.

In fact, an early move towards presenting the CASE metadata to the Oracle user community was achieved by carefully exporting the Systems Engineer model (held in the Gupta SQL*base rdbms) straight into Oracle itself. The initiative was successful, although suffered from the problems of maintaining currency between the two environments.

The continuing development of this work, extending and populating the Data Architecture over several years, resulted in the emergence of a new and profound data management problem which had to be addressed if we were to continue to obtain benefit from our use of Oracle. Simply, our schema was becoming overly complex for the environment in which it was being described. A state of ‘data overload’ was fast approaching.

The Problem Defined

The problem of schema complexity became more acute as additional models were migrated into Oracle, simply because the highest level of object abstraction within Oracle is the table (or in special cases the view). The BGS view of data demanded a means of aggregating suites of tables into logical groupings which related directly back to the source of the data, e.g. the geological discipline from which the data was derived, simply as a means of distinguishing discrete data components from within what was becoming an increasingly complex heterogeneous assemblage of database objects. For example, there are of the order of 4400 tables across the BGS Oracle system. It was important to obtain this aggregation within Oracle itself, because so many staff were now (correctly) regarding Oracle as a comprehensive repository. Staff were becoming used to the concept of both data, and 'data about data' all being held within the same environment. Further, there was no possibility of embedding this sort of technical metadata into the application side (which would be a suspect design methodology in any case) because there is too wide a variation in application methods employed across the Survey.

The Solution

The solution to the problem of schema complexity lay in the arbitrary definition of the concept of 'the databank' – which BGS has defined as '*a discrete aggregate of Oracle tables and associated database objects which collectively comprise a meaningful data set*'. By database objects, we include views, indexes, synonyms, triggers, constraints, etc. Databanks are reflections of distinct geoscience collections, for example the 'borehole databank', or the 'geochemical databank', and so on. Currently, there are almost 90 databanks within the BGS schema.

A suite of additional tables has been defined and established to hold information within Oracle to describe databanks, and their relationship to the existing Oracle system. These tables are referred to as **technical metadata**, and have been tightly integrated into the Oracle data dictionary system itself. They comprise the 'technical metadata databank', to maintain symmetry with the principle they support. We have been scrupulous in not modifying in any way the end user view of the conventional Oracle data dictionaries, but rather have extended them by using additional tables.

In the first instance the new table 'meta_dbank' contains the name, the manager, and a description of each databank. Implicit in this definition is the requirement that every time a new databank is defined a case for its creation has to be made to the BGS database administrator, it must have a member of staff nominated to take responsibility for it, and that databank must be described in a simple textual manner, so that any person browsing the database for databank descriptions would understand immediately what the nature of the data was that was referenced by that databank.

Consistent with our policy of not modifying, but rather adding separately to the Oracle data dictionary, we defined a table called 'meta_objects', which effectively maps as a one-to-one relationship to the Oracle dd view 'dba_objects'. Only the attribute 'status' was carried over to 'meta_objects' (to save having to make a join with dba_objects). The other attributes in 'meta_objects' were required to support the metadata model. These included in particular an object description, and crucially two responsibility fields. The **data administrator** is that person who has responsibility for the intellectual content of the table, who best understands the actual data. The **applications manager** is that person charged with responsibility for the design of the table (and its associated objects), and for building the applications to support the query and maintenance of the data in the table.

These extensions are completed by a third table 'meta_dbank_objects', which functions as an associative, or link table between meta_dbank and meta_objects, thus resolving the many-to-many relationship between meta_objects and meta_dbank_objects. This is to support the model which states that a database object may be a member of more than one databank.

The most simple example of this many-to-many relationship is to consider the large number of geoscientific dictionary tables which are used throughout the BGS. These are codes, grouped by domain, and stored in conventional, individual Oracle tables, which are used across the entire system. Their classification has resulted in a suite of databanks, of which two examples are the 'geoscientific codes' and 'administrative codes' databanks. The dictionary tables present in these databanks are obviously to be found in several other databanks, in support of code expansion used by their various constituent tables. So, for example, the 'borehole databank' would consist of tables used to store primary data, pertaining directly to borehole material, samples, documents etc, as well as the numerous dictionary tables used in their support. Many of these same dictionary tables would also be cited in several other databanks.

Figure 1 illustrates, in a schematic form, how the relationship between a part of the BGS technical metadata and the relevant parts of the Oracle data dictionary may be expressed in a simple ER diagram.

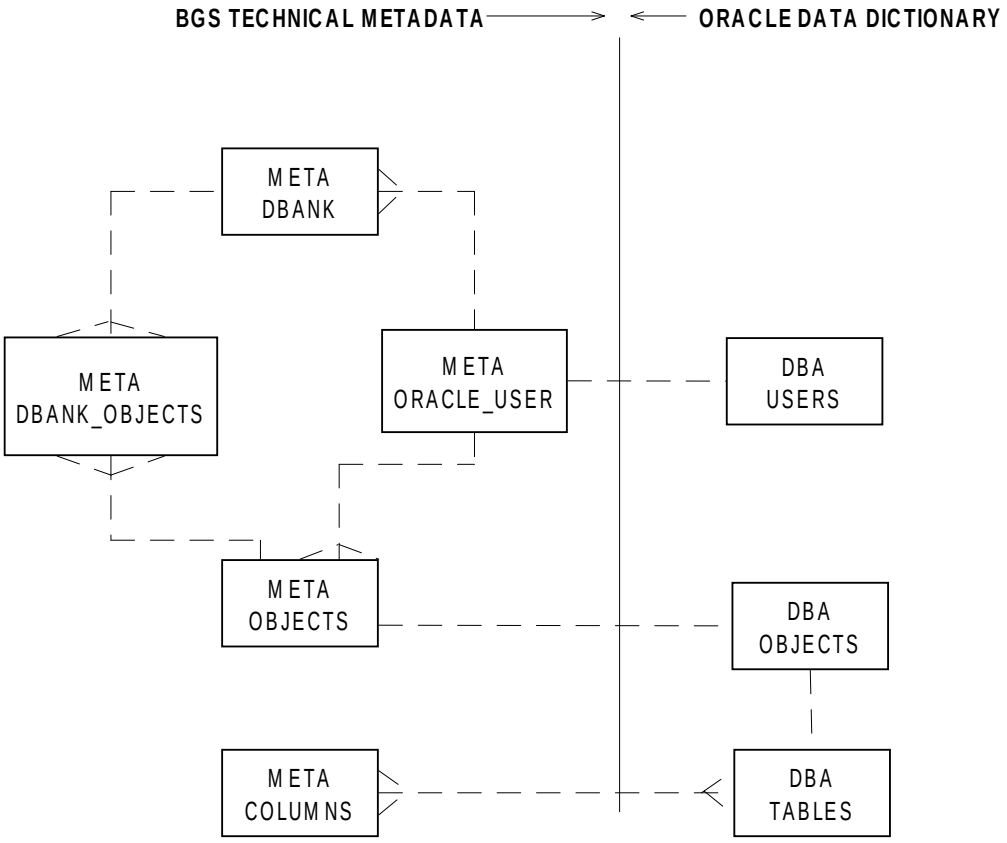


Figure 1.

A combination of triggers and constraints is used to ensure the integrity of these technical metadata tables. Thus no objects can be inserted into 'meta_objects' unless they are already present in 'dba_objects'. No objects can be inserted in the associative table 'meta_dbank_objects' unless that object is already present in 'meta_objects'. Similarly, only databanks already declared in 'meta_dbank' can partake of an associative relationship in 'meta_dbank_objects'.

Additional metadata tables have been defined as well as the three major ones discussed above. For example, 'meta_oracle_users' enables us to add information pertaining to Oracle users, including an expansion of their username into their full name. We have also started to construct a metadata table to enable us to document more fully important attributes (described as 'crucial

database fields') which are vital to the developing standards emerging from the metadata. Again, this table is based on an extension of dba_tab_columns, and is consistent with it.

It will be noted that the Oracle system table 'dba_objects' has been mentioned. This raises a point concerning privileges and access across Oracle. Consistent with our policy of subsuming all corporate data into one schema, it was regarded as essential that all staff should be able to see all data across the entire Oracle system. We start from the standpoint that all data is owned by the BGS, all staff are employees of the BGS, and therefore, unless there are overriding commercial reasons (and there are in a limited number of such cases), the entire data resource should be available for all staff. This is an attempt to encourage synergy, to look for the added value benefit of seeing data – not as individual islands of data, but rather as part of a larger, corporate whole. There is no doubt that this approach is paying off.

Therefore, to ensure that all staff can have, at the very least, select access across the entire Oracle system, several of the underlying Oracle data dictionary views (such as dba_objects, dba_tables, dba_triggers, dba_constraints, etc) have been opened to public access (where 'public' equates to the BGS community). This has been done, not by granting select on these objects, but rather by creating a suite of views for each of the underlying data dictionary views, and then granting public select on these. This provides a consistency for staff, where the corporate schema itself is known as 'bgs', and therefore they become used to the prefix 'bgs' as indicating a corporate object.

The following figure illustrates the relationship between the various levels of database access, and the way in which we have created the corporate views to by-pass the restrictions of the default 'user_%' and 'all_%' object views:

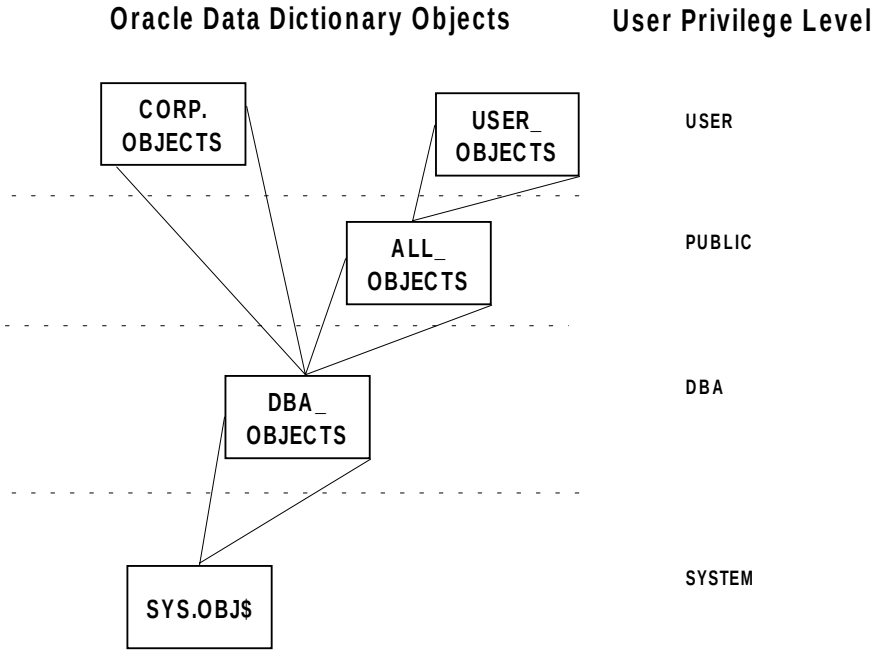


Figure 2.

Why Technical Metadata is important

A discussion of this topic has already been inferred, in the sense that the superimposition of a logical structure onto a large undisciplined collection of tables was already felt to be necessary.

However, the overriding reason why technical metadata is so important is that it fundamentally underwrites the quality of the data being managed. Data documentation, management responsibilities and procedures, and business rules are all intimately bound up with data quality, and it is these crucial parameters that are held in the technical metadata.

Once this infrastructure was in place, it became much easier for three particular groups of people to make use of the coporate data resource.

Firstly, staff responsible for discrete sets of data were able to review whether overlap existed between their data sets, and those held in other parts of the organisation. This has provided several examples of instances where duplicate data structures holding structurally very similar data have been assimilated into one uniform structure.

Secondly, a major part of the Survey's work is responding to requests from the public at large for geoscience information. It is now very much easier for BGS enquiry staff to answer such questions, often during the telephone conversation from which the request is coming. Previously, these requests could only be met by writing a letter to the enquirer, after a time-consuming search of the various paper data repositories, or running lengthy batch queries.

Thirdly, application developers, whether building new applications, modifying existing applications, or building new Oracle tables, are now able to review the metadata, better to maintain consistency with already existing standards. This same approach has been used, admittedly in a somewhat empirical fashion, to derive and establish corporate standards for all application developments.

As the metadata has grown, so diversity from single standards have been thrown up. Undesirable though this diversity may be, the first step to tackling it is to document it, and this has unquestionably been one of the benefits of our emphasis on metadata. Then, through a series of negotiations, staff have moved increasingly towards single standards. In fact, it was quite revealing how much positive support there has been from staff for the convergence towards single standards. Standards are recognised as the very basis of successful scientific work.

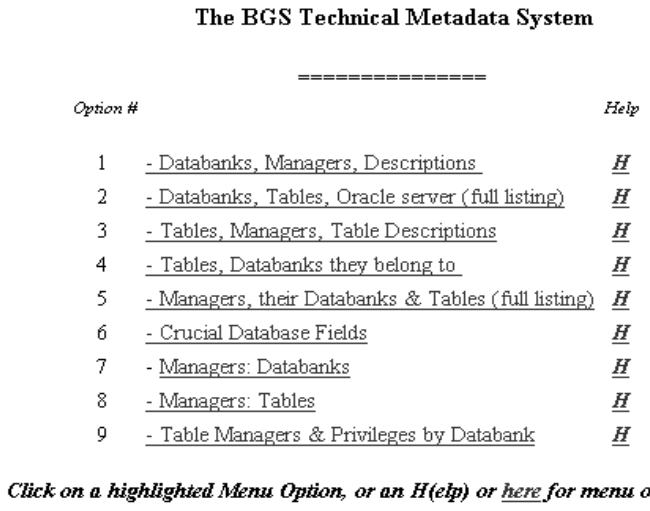
It will be realised that the imposition of standards in a didactic manner across the Survey has not been a course we have followed. Quite simply, given the nature of the organisation, this approach would not have worked. The adoption of corporate standards, and their ready availability, has been something of a cultural issue in the BGS, and progress in this direction has had to be made in a distinctly diplomatic fashion. We were warned in 1990 by Logica, that the adoption of a corporate Data Architecture, and all that flowed from it, was as much a cultural challenge as a technical one, and this has proved very much to be the case.

User Access to the Technical Metadata

It was always regarded as important that all staff would have easy access to the technical metadata. This has been provided through intranet pages for some time now, using report writing facilities, which export an html version of the various technical metadata reports which have been built.

This approach had two benefits. As one of the earliest uses for the BGS intranet, it introduced staff to the importance of the intranet as a means of disseminating crucial information. Secondly, it enabled us to retain all information within Oracle itself, and (as referred to previously) to drive Oracle hard as a total repository solution, rather than just a place to store prime data. The incorporation of a mature technical metadata as part of the relational database environment itself is a key component of our strategy.

Report pages have been constructed which list a variety of combinations of metadata components one against the other. The figure below is the intranet menu page which allows the display of these various reports. Of course, the competent user of sql is able to interrogate the database directly without having to go through the intranet route, if so required.



The above menu betrays a slightly more complex situation than has actually been described (ref option 2). In fact, our Oracle environment is spread across two Oracle instances, geographically remote from one another, and connected by SQL*Net over a WAN. To the end user this is irrelevant, as the same metadata construct exists on each instance.

Conclusion

The introduction of the concept of the databank has been successful, and continues to be so. It has strengthened and endorsed the importance of technical metadata. It has facilitated rapid, ad hoc queries of the database. It has aided greatly our development of corporate-wide standards, and has consequently assisted in the cultural move towards the perception of data as a corporate resource. To any organisation with a similarly heterogeneous assemblage of data, we would recommend the concept of the databank as a means of obtaining greater value from their use of the Oracle rdbms.