Geochemistry Data Model Summary

For the Open Geoscience Data Models Project

Carl Watson & Iestyn Evans

BGS Database Solutions Team

Contributions from:

Martin Nayembil, Alan MacKenzie, Susan Hobbs, Neil Breward, Christopher Johnson, Robert Lister, Linda Ault

We believe that most readers will imagine the data model described in this document as a physical database, therefore we use the terms 'Table' and 'Column' where 'Entity' and 'Attribute' would be more technically correct.

1. Introduction

1.1. Document purpose

This document provides a summary of the Geochemistry Data Model produced by the BGS for the Open Geoscience Data Models project. It includes descriptions of the main components (also known as tables or entities) which make up the data model, the relationships between these components and the dictionaries which are, in effect, controlled vocabularies of the terms used in describing, and supporting the description of, scientific and other observations.

This document does not deal with technical details relating to specific database hardware or software, for those please see the implementation folder in the download**.**

1.2. Simplified data model

In this download we have provided a simplified data model based upon the current BGS Geochemistry Database. This database holds the results of the chemical analysis of samples from the land area of Great Britain and Northern Ireland (excluding groundwater). Relevant information on the method of analysis, limits of determination (detection limit, upper concentration limit) and sample type are included.

We have simplified the model because we want to highlight the essence of the database but not to present an overly complex design. We have removed many features and tables which relate to concepts which are specific to the BGS but which may not be relevant to other organisations.

We hope that this design can provide one example of how you might structure a Geochemistry data model while highlighting areas that require further consideration when designing a geochemical database.

1.3. BGS standards and naming conventions

The creation of the model has been carried out by BGS staff using corporate database standards and best practices.

The table and column names have been altered to more generic versions of those in use on the BGS corporate database, however, the names used in this data model conform to the BGS database naming conventions, for example all dictionary names should begin with DIC_.

1.4. A brief history of the BGS Geochemistry Database

The following bullet points are based upon BGS Reports and discussions with current BGS staff:

- The BGS designed a relational database for geochemical results in 1986 as part of the Mineral Reconnaissance Programme (MRP). This database was the starting point for the geochemistry database that we use today.
- Care was taken to ensure that the initial design would allow the capture of data from the main geochemical programmes from that time – the MRP, Geochemical Survey Programme (GSP), Mineralogy and Petrology and Biostratigraphy.
- This first geochemistry database was implemented on an Oracle 6 server.
- During 1988 a 'menu-driven' front end was developed for theMRP database that would for the first time allow non-expert users to access the data.
- In 2002 it became standard practice to populate a data qualifier field in order to convey more information about data quality to data users, e.g. concentration below detection limit or measured concentration less precise due to the presence of high concentrations of interfering elements. This has proved particularly useful when data have been extracted for use by external customers.
- Over the years there have been regular updates made to the design to add new or alter existing tables and columns to capture new information as well as to standardise the design to enable greater links with other databases.
- By the beginning of 2012 there were more the 10.5 million analyte determinations (measured chemical quantities) recorded in the database.

The Geochemistry Database is actively populated with geochemical data from current projects generating geochemical data from the UK land area together with historical datasets (mainly MRP data) in paper and digital formats. The majority of new data are field data and sample analyses from continuing strategic surveys such as the Geochemical Baseline Survey of the Environment (G-BASE – formerly GSP) (http://www.bgs.ac.uk/gbase/).

For more information on the history of the BGS Geochemistry Database see these publications:

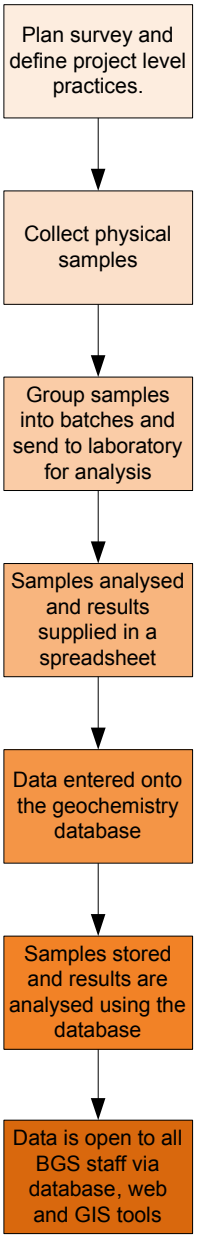
- **COATS, J S.** *2004. The BGS Geochemistry Database: history, design and current usage. British Geological Survey, Internal Report Series IR/04/033.*
- **HARRIS, J R, and COATS, J S.** *1992. Geochemistry database: data analysis and proposed design.* British Geological Survey Technical Report*, WF/92/5.*
- **MacKenzie, A.C.**; **Johnson, C.C.**. 2006 *Loading analyte results to the Geochemistry Database using the MS ACCESS LOADER application.* Nottingham, UK, British Geological Survey, 45pp. (IR/06/097)
  http://nora.nerc.ac.uk/7400/

1.5. The BGS geochemistry workflow

Before we look at the details of the data model this section will briefly describe a typical workflow that would result in data being entered into a geochemistry database.

Within the BGS almost all work is project based and the following workflow steps describe how a fictional geochemical survey project might proceed:

- A survey plan is created; it includes details on sampling locations and the types of material to be collected.
- Procedures are defined for the control of sample collection, sample labelling, and preparation for laboratory analysis.
    - Samples are collected, numbered and bagged up as per the procedures. For example, for G-BASE the samples and the sites from which they are collected are described with the unique sample id and grid coordinates. These form a simplified 'field database'. For more information see Lister et al., 2005. The G-BASE field database.
- Samples are grouped together in batches relevant to the needs of the project. There could be one batch for a project, for an area or time period.
- The sample batches are prepared and stored for analysis as per the defined procedures.
- The samples are analysed, project requirements determining which element concentrations are measured. Analysis may be by a variety of techniques, dependant on the species to be determined.
- The results of the analyses are referred to as the laboratory data, commonly supplied in spreadsheet format for entry into the database.
- Batches of samples are moved to the long-term storage facility once analysis is complete.
- Once entered into a geochemistry database, the results are available to all BGS staff and can potentially be exposed to the public via the Open Geoscience website.
    - BGS staff can use a variety of tools to connect to the results in the database including tools such as TOAD and SQL Developer, Microsoft Access and ArcGIS.
    - We have a business rule that no analytical information is accepted in the BGS Geochemistry Database unless there is site coordinate information



For more information on the workflows for the population of the BGS Geochemistry Database please refer to the following publications, although it is worth noting these focus on a specific project called G-Base and may contain practices which are not typical:

- **Ault, L.**; **MacKenzie, A.C.**. 2006 *From LIMS to geochemistry database: GBASE samples analytical data.* Nottingham, UK, British Geological Survey, 35pp. (IR/06/075) http://nora.nerc.ac.uk/7299/
- **Lister, T.R.; Johnson, C.C.**. 2005 *G-BASE data conditioning procedures for stream sediment and soil chemical analyses.* British Geological Survey, 85pp. (IR/05/150) http://nora.nerc.ac.uk/11170/
- **Lister, T.R.; Flight, D.M.A.; Brown, S.E.; Johnson, C.C.; MacKenzie, A.C.**. 2005 The G-BASE field database. British Geological Survey, 84pp. (IR/05/001) http://nora.nerc.ac.uk/8756/
- **Johnson, Christopher. 2005** G-BASE field procedures manual. British Geological Survey, 65pp. (IR/05/097) http://nora.nerc.ac.uk/5190/

2. The data models explained

2.1. Current BGS Geochemistry database

As mentioned earlier the current BGS Geochemistry Database has evolved over many years and has become quite complex. It will not be described in detail in this document but the complexity can be seen in the following diagram and subsequent information.

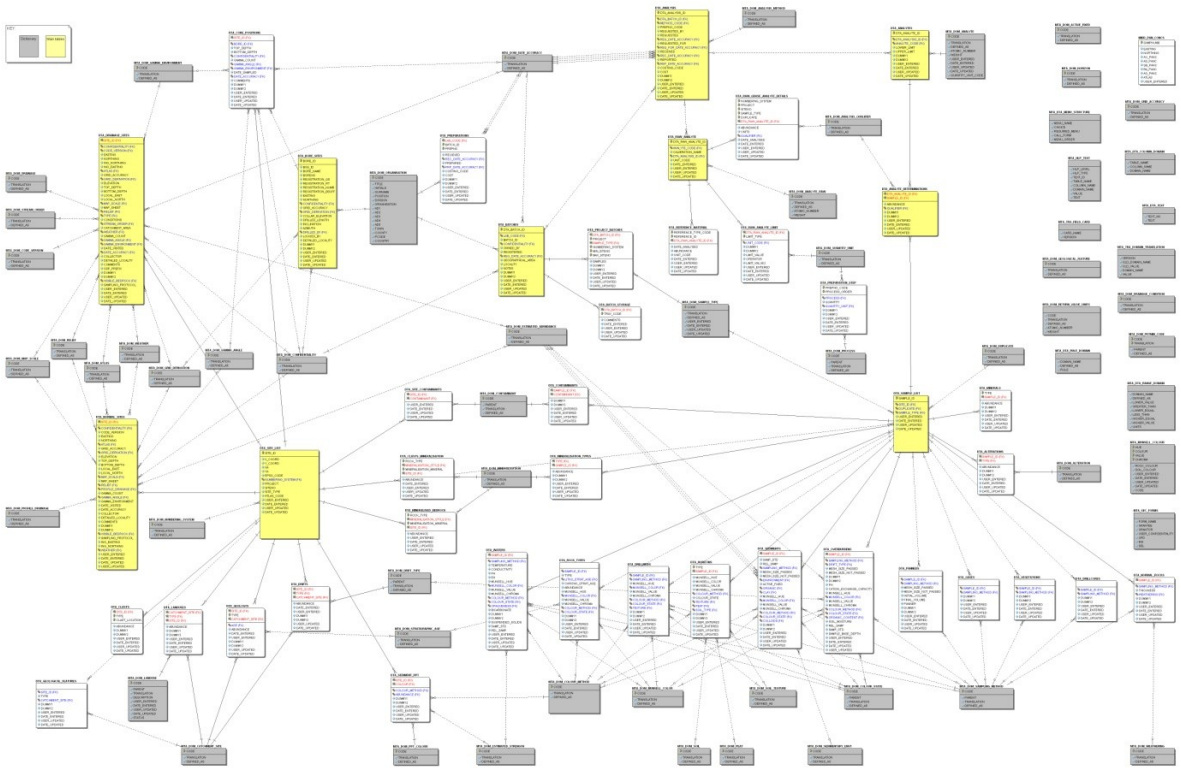


Figure 1

- The main tables (recording location, sample-batch relationships and analyte determination details) are shown in yellow.
- The white tables contain additional information for specific situations and are not universally populated. For example, the sample table (one of the main tables) contains columns for values that apply to all samples but it is linked to 12 or so additional tables that are used to store details specific to each type.
- The dictionaries are shown in grey.

There are a group of dictionary tables on the right hand side unconnected to the other tables, these are the result of a historical data entry system which has been replaced and these tables will be removed in future updates to the database.

For more information about the BGS Geochemistry Database and its uses there are a range of publications available online or via the NERC Library. Some are listed in the section - A brief history of the BGS Geochemistry database.

2.2. The Geochemistry Data Model for the Open Geoscience Data Models project

For more technical details and a more complete "Entity-Relationship" Diagram see the technical documentation that accompanies the implementation code.

The following diagram shows the new and simplified logical design which has been created as part of the Open Geoscience Data Models project:

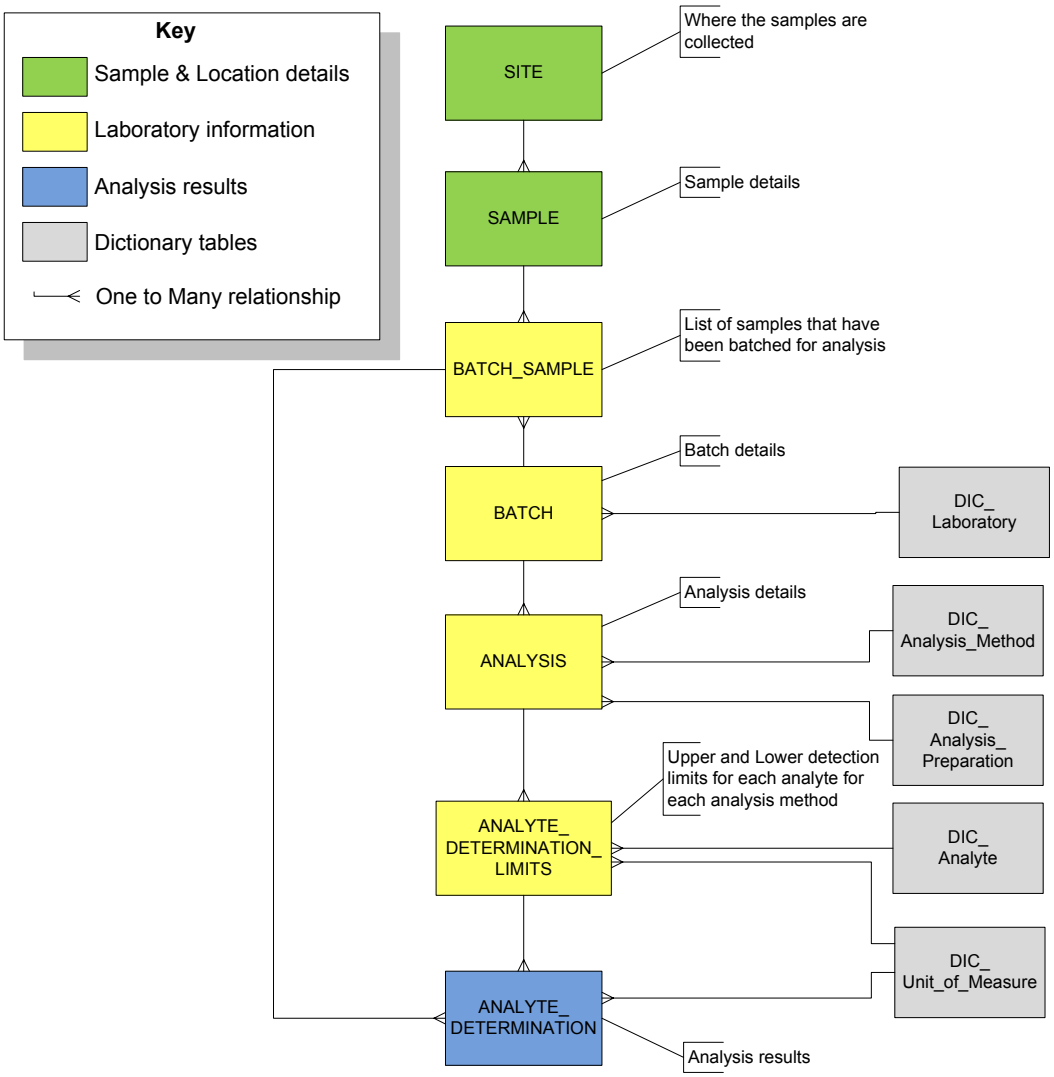


Figure 2

2.3. Components of the Data Model

The data model is composed of four types of tables, they are:

i. Sample & Location details (sample and field site specific information)
ii. Laboratory information (batches and analysis methods)
iii. Analysis results
iv. Dictionaries (controlled vocabularies) used to constrain the data, i.e. restrict the data that may be entered for certain values.

It is standard practice in BGS corporate database design to assign a three letter prefix to table names to indicate the general subject areas to which these tables belong. In this case we have used the prefix 'CHM' to represent Chemistry.

Site Information Table: CHM_SITE

This table is used to record the sampling location. No sample should be accepted into the database without site coordinates. One or more samples could be collected from a single location.

Typical information to be captured in this table:

- Coordinates
- Coordinate reference system
- Type of location, for example – drainage sites, borehole or normal

Sample Information Table: CHM_SAMPLE

This table holds the main descriptive details for each sample.

Each sample comes from a single location as indicated by the link to CHM_SITE and is usually batched for analysis reasons, hence the link with CHM_BATCH_SAMPLE, the link is a one-to-many relationship because it is possible for a sample to be sent for analysis in more than one batch, but this is rare.

Typical information includes:

- Collection date
- Who collected the sample
- Sample material type, for example – Rock, Gases, Mud, Waters
- Properties such as size, colour and mesh size

Sample Batch Identifier Table: CHM_BATCH_SAMPLE

This is a linking table that is used to record which samples were batched together for the purposes of analysis.

In most cases it will contain details of many samples being part of a single batch, however, it is possible to send a single sample to more than one laboratory for analysis, and therefore the design of this table allows a sample to be recorded in several batches.

This table contains a column for the identifying field in CHM_SAMPLE (sample_id) and another for the identifying field in CHM_BATCH (batch_id). In addition there is an optional laboratory code column, this would become mandatory if you used different laboratories who could

Sample_ID	Batch_ID	Laboratory_Code
1	A	NCL
2	A	NCL
3	B	NCL
4	B	NCL
5	A	CIL

supply the same batch_id for different batches.

Batch Information Table: CHM_BATCH

This is an index table for recording details relevant to a batch of samples such as the laboratory a batch is sent to, reason for analysis and any other information that should be directly associated with a batch. Each batch could be subjected to one or more analysis methods and this information is captured through the link to the CHM_ANALYSIS table.

Analysis Information Table: CHM_ANLAYSIS

This table contains the key information detailing the methods of analysis a particular batch has been subjected to and what preparation methods were used.

For each type of analysis, represented as a record in the table CHM_ANALYSIS, there will be one or more analytes which could be determined, hence to the link to CHM_ANALYTE_DETERMINATION_LIMITS.

CHM_ANLAYSIS_DETERMINATION_LIMITS

In this table you record what the detection limits are (upper and lower) for each analyte for each method of analysis listed in CHM_ANALYSIS.

There will be one row in the table for each analyte and method used, for example the table below shows how copper limits may be recorded for a couple of analysis methods.

Analyte_Code	Method_Code	Lower_Limit	Upper_Limit
Cu	XRF	10	10000
Cu	ICP	5	12000

Analytical Results Table: CHM_ANLAYTE_DETERMINATION

This is the table which contains the results of the geochemical analysis on a batch of samples. It captures the abundance of analytes for each method of analysis used to test a particular sample.

We chose to link this table to the batch and sample table CHM_BATCH_SAMPLE but it would also be valid to link directly to CHM_SAMPLE.

3. Points of interest and suggested changes

This section contains a few points we wish to highlight as we believe they could help anyone wishing to use the data model presented earlier in a real world situation.

The data model we have presented is only a skeleton and requires alterations, additional columns and tables to suit local or organisation specific requirements, the following comments are intended to stimulate thoughts on what those changes could involve.

Sample Sites / Locations

- In the event of collecting a sample there is often a one to one relationship between the location and local weather conditions at the time, therefore it is possible to use a single table to capture all of this information. However, if you wanted to take multiple samples from a single location it may be more efficient to reuse locations and capture weather details separately.
    - Depending on which approach you choose to take will affect where you record information such as:
        - sample collection date/time
        - local weather conditions at time sample of collection
        - any other temporal conditions which might affect the chemistry of a sample
- When putting together the Geochemistry data model for the Open Geoscience Data Model download we spent some time debating how to structure the tables (normalise the entities) that describe the location of a sample site.
    - You are likely to want to identify the types of location from which a sample was collected, for example, in the BGS we have three locations types:
        - Drainage site (Stream, Lake etc)
        - Borehole
        - Normal site (Any other site that is not a drainage site or a borehole)
    - You could do this by simply adding columns to the CHM_SITE table to hold site type specific details OR add supporting tables to capture details unique to each site type and link these back to the main site table.
- You may wish to link into existing tables on your database that contain information about a location such as field survey sites or boreholes. For example the CHM_SITE table could contain a column for BOREHOLE_ID that linked to a Borehole Index table such as that available in the Borehole data model available through the Open Geoscience Data Models project as shown in Figure 3.

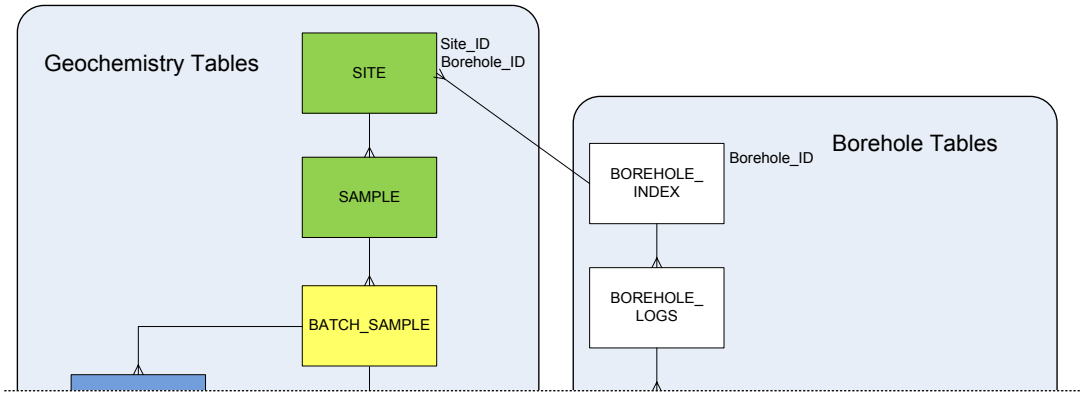


Figure 3

Project / survey specific details

Within the BGS all of our geochemical datasets are associated with a specific project. When we populate our database with analytical results a project code is recorded so that we can link the results to project specific information, for example sample collection protocols documented as part of the project planning phase.

- If you do not associate samples with a project or similar equivalent you could capture the information such as collection protocols directly in a geochemistry database. The best places for this information in our geochemistry data model would probably be in the CHM_SITE or CHM_SAMPLE table.

Sample to batch relationships

Should you explicitly or implicitly store which samples are included in each batch?

- In the BGS Geochemistry Database you can only deduce which samples were in a particular batch by looking in the CHM_ANALYTE_DETERMINATION table, there is no table to explicitly capture these links. This approach is suitable for most of our requirements and fits in with traditional workflows and requires one less table, however there are some possible limitations.
    - For example; if a sample has been allocated to a batch, but has not been analysed yet, it will not be present in the CHM_ANALYTE_DETERMINATION table and therefore it is not possible to identify which batch it was a part of. This could be a problem if one of the samples in a batch were to be lost, it would not be possible from the database to identify what batch and therefore laboratory it had been assigned to.
- Alternatively, in the geochemistry data model for the Open Geoscience Data Models project we decided to add the table CHM_BATCH_SAMPLE to explicitly capture sample to batch relationships. With this design it is possible to populate the database with site, sample and batch details before analysis results had been received.

- To a certain extent it depends on your workflow and whether you want to record sample and batch details before the analysis results are returned. If you need to record the sample and batch details up front you need an equivalent to CHM_BATCH_SAMPLE, if not you could use a design like that shown in Figure 4.

The following diagram (Figure 4) shows how the data model could be re-structured to capture batch information implicitly.

- Advantage – simplified design by removing CHM_BATCH_SAMPLE
- Disadvantage – possible loss of sample to batch data. In the diagram below we have shown how sample C which was batched into batch X appears to have never been batched simply because there is entry for it in the analyte_determination table.

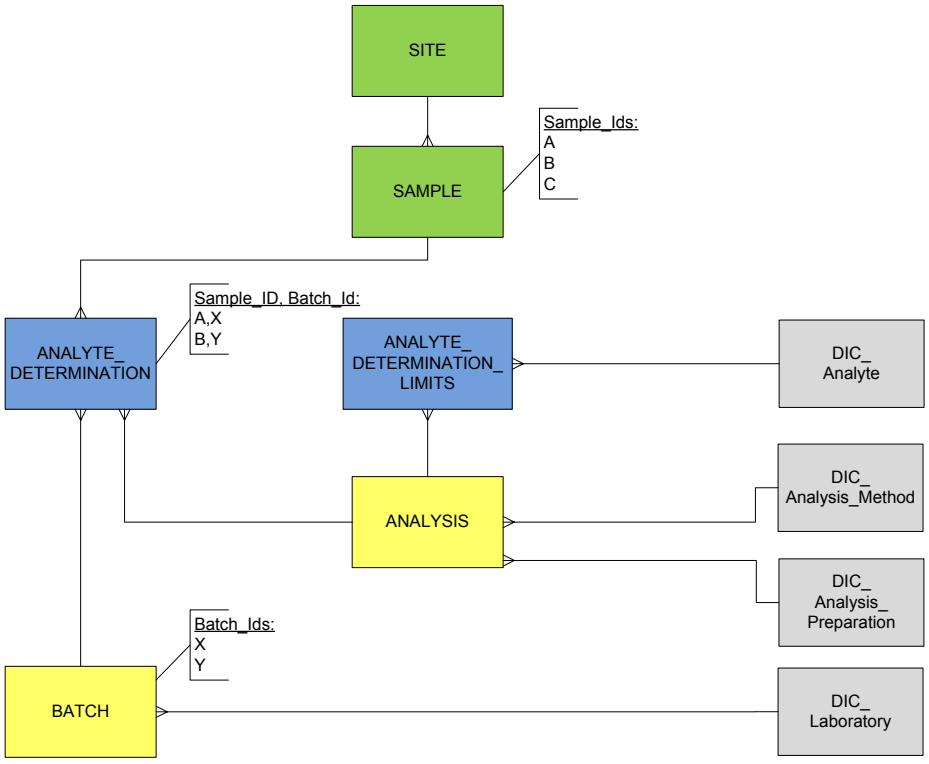


Figure 4

Dictionaries

Dictionaries are, in effect, controlled vocabularies, they control the terms we wish to use in describing, and supporting the description of, scientific and other observations.

- We strongly recommend using controlled dictionaries for the following concepts:
    - Analysis Methods
    - Analysis Preparation
    - Analytes
    - Laboratory
- You may also want to add dictionaries to describe the following sample and sampling site features:
    - Weather conditions at the sample site at the time of collection
    - The stream flow at the sample site at the time of collection
    - Sample contamination details
    - The presence or absence of rock outcrop

Example of a Dictionary for weathering:

CODE (Unique ID in a readable form)	**TRANSLATION** (Short description, used in lists etc)	**DESCRIPTION** (Full sized description)
NEW	Fresh	No visible sign of rock material weathering (eg drill core sample).
MNR	Faintly weathered	Discoloration on major discontinuity surfaces.
SLT	Slightly weathered	Discoloration indicates weathering of rock material and discontinuity surfaces. all the material may be discoloured by weathering and may be somewhat weaker than in its fresh condition.
MOD	Moderately weathered	Less than half of the rock material is decomposed and or disintegrated to a soil. Fresh or discoloured rock is present either as a discontinuous framework or corestones.
HGH	Highly weathered	More than half of the rock material is decomposed and or disintegrated to a soil. Fresh or discoloured rock is present either as a discontinuous framework or corestones.
CMP	Completely weathered	All rock material is decomposed and or disintegrated to soil. The original mass structure is still largely intact.
TOT	Residual soil	All rock material is converted to soil. the mass structure and material fabric are destroyed. There is a large change in volume, but the soil has not been significantly transported.

Sample numbering / referencing

You need to consider what sample identification systems to use and labels to assign to the physical sample when it is collected in the field. Consider how these might be used to locate the sample once it has been entered into long-term storage.

- This is likely to vary for most organisations.
- In the BGS we have a large number of historical samples collected over the decades, these have traditionally been identified by project code + an incrementing sample number.
    - The project code and sample number recorded in our database are the same as those recorded on the physical field report and the sample itself.

        - For the purposes of quality control the sample number may be in random rather than incremental number order
    - Sample numbers are linked to Batch numbers in one of our database tables and our physical batches put into numbered trays, these trays are arranged in numerical order in the storage facility. It is also a required protocol that samples are analysed in numerical order.
        - We have not included tray locations in the data model but we strongly recommend that you add a table to store details of where samples or batches are physically located.

Should the sample number / reference contain additional information?

- Some organisations like to embed additional information into identifying fields, others prefer to keep identification and additional metadata separate:
    - As mentioned above, the BGS have embedded project codes into the sample identifier but this could and possibly should have been captured as a separate Project_ID
    - You may also incorporate information about sub samples and duplicates as described in the following section.

Sample duplicates & subsamples

We store details on the following sample types:

- One off / original samples
- Subsamples, where a new sample is split into smaller samples
- Duplicates, identical samples

How should you manage the relationships between samples, duplicates and subsamples?

It is possible to devise naming conventions that mean the structure of the sample ID could be used to identify links between samples, for example see Figure 5 below we use the convention:

- "SMP_" equates to original sample
- "SUB_" indicates subsample
- "DUP_" indicates duplicate

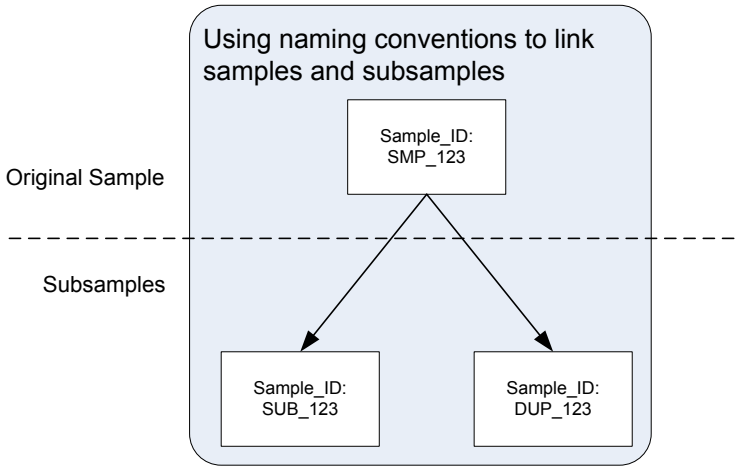


Figure 5

An alternative methods for managing subsamples could be to add a Parent_Sample_ID attribute in the CHM_SAMPLE table so that any derived sub samples could be tracked back to the parent material as shown below (Figure 6).

In both approaches it is assumed that the samples are held in the one table CHM_SAMPLE.

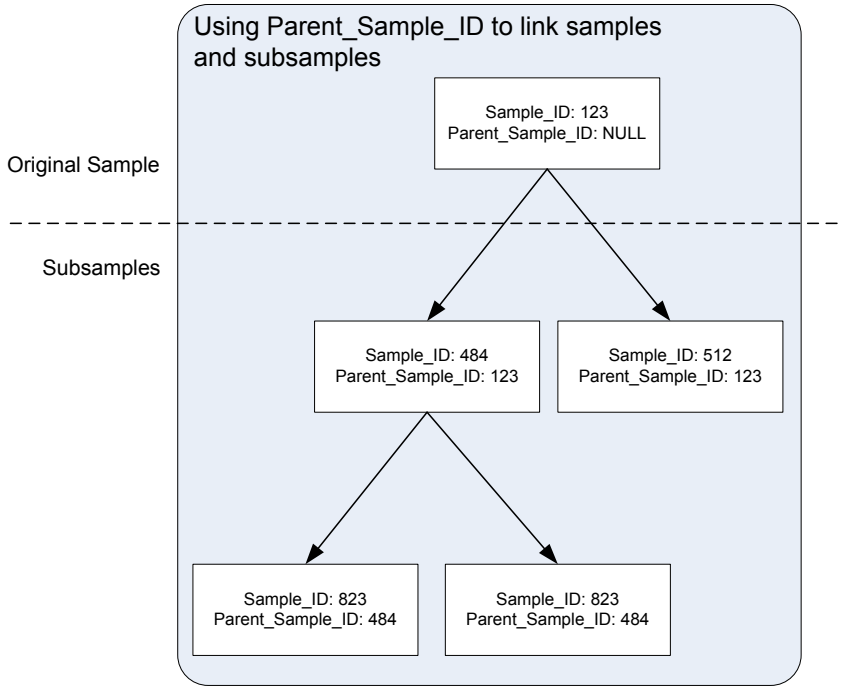


Figure 6

Do you wish to store different versions of the analysis results?

It is necessary to 'level' the raw data in order to combine the analytical results for a given batch of samples with those subjected to different analysis methods to create a geochemical database with a national or regional scope. For a detailed example of how geochemical data has been levelled by the BGS see Johnson, C *et al* (2008).

> **Johnson, Christopher; Ander, Louise; Lister, Bob; Flight, Dee.** 2008 Data conditioning of environmental geochemical data : quality control procedures used in the British Geological Survey's regional geochemical mapping project. In: de Vivo, B.; Belkin, H.E.; Lima, A., (eds.) *Environmental geochemistry : site characterization, data analysis and case histories.* Amsterdam; London, Elsevier, 93-118. (IR/05/097)
>
> http://nora.nerc.ac.uk/3722/

It is possible to capture the raw data (laboratory results), levelled analysis results or both in a single data store. If you do choose to capture multiple versions of analysis results in a single database you need to ensure that the database design allows you to capture the provenance for each version.

- In the BGS Geochemistry data model we have both raw and derived data, however we decided to keep this Geochemistry Data Model simple and chose not to include separate structures for raw data capture.
- Options include:
    - Add separate analyte determination tables, i.e. in the current BGS Geochemistry database we have a RAW_ANALYTE_DETERMINATIONS table and the ANALYTE_DETERMINATIONS table which contains the levelled values
    - Add an ANALYTE_DETERMINATION_TYPE field to the CHM_ANALYTE_DETERMINATIONS table, where a type could include 'raw data' or 'levelled'.
    - Or simply have a clearly defined business rule that only raw or levelled data was to be entered into the database.

Analyte Determination Units of Measure

A decision was taken a few years ago that the BGS Geochemistry Database should be standardised to store all analyte concentrations in ppm (parts per million or mg/kg). The aim of this standardisation was to avoid confusion by mixing different units of measurement. Previously 'major' elements were usually shown as 'weight percent' of the oxide, for example CaO %. You can see where there may have been some confusion, as elements such as manganese were sometimes shown in ppm, and sometimes as MnO %. If you wish to capture values using different units this could be achieved by adding a UNITS column to the CHM_ANALTYE_DETERMINATION

table, if you choose to do this it would be wise to constrain the units through the use of a UNITS_OF_MEASURE dictionary.

- Whichever approach you choose to use, the upper and lower detection limits that are captured in the CHM_ANALYTE_DETERMINATION_LIMITS table should be recorded using the same units as those in the CHM_ANALTYE_DETERMINATION table.
- It is also good practice to also hold the original units for reference, provenance, conversion to a different unit (it's always better to convert from original units than a converted unit to reduce the potential error multiplying effect) or for data correction issues on the standardised units.