Geochemistry Geophysics Geosystems G3
AN ELECTRONIC JOURNAL OF THE EARTH SCIENCES
PUBLISHED BY AGU AND THE GEOCHEMICAL SOCIETY

Technical Brief
Volume 1
May 24, 2000
Paper number 1999GC000026
ISSN: 1525-2027

A global geochemical database structure for rocks

K. Lehnert, Y. Su, and C. H. Langmuir
Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964
(lehnert@ldeo.columbia.edu; ysu@ldeo.columbia.edu; langmuir@ldeo.columbia.edu)

B. Sarbas and U. Nohl
Max-Planck-Institut für Chemie, Abteilung Geochemie, Postfach 3060, Mainz, D-55020 Germany
(sarbas@mpch-mainz.mpg.de; nohl@ldeo.columbia.edu)

[1] **Abstract:** This technical brief describes a geochemical and petrological database structure based on the relational model that has broad applicability to chemical analyses of geological materials. Notable features of the database structure are its comprehensiveness and flexibility. The structure consists of 34 interrelated tables, which can accommodate any type of analytical values for all different materials of rock samples (volcanic glasses, minerals, inclusions, etc.) and for samples from any tectonic setting. A broad spectrum of supplementary information (metadata) is included that describes the quality of the analytical data and sample characteristics, such as petrography, geographical location, and sampling process, and that can be used to evaluate, filter, and sort the chemical data. All data in the database are linked to their original reference. The database structure can be implemented in any relational database management system (RDBMS). It is currently applied in two different rock database projects (RidgePetDB and GEOROC).

Keywords: Database; petrology; geochemistry; data management.

Index terms: Geochemistry; instruments and techniques.

Received October 27, 1999; **Revised** December 8, 1999; **Accepted** December 9, 1999; **Published** May 24, 2000.

Lehnert, K., Y. Su, C. H. Langmuir, B. Sarbas, and U. Nohl, 2000. A global geochemical database structure for rocks, *Geochem. Geophys. Geosyst.*, vol. 1, Paper number 1999GC000026 [6408 words, 3 figures, 34 subfigures, 3 tables]. May 24, 2000.

1. Introduction

[2] Development and testing of models for the Earth's dynamic systems require the scientific community to take full advantage of all data available. In the past this has been attempted by individuals making spreadsheets on their own computer for the particular problem at hand. However, the amount of geochemical data produced and published in the Earth sciences is growing exponentially. Twenty years ago a dozen elements on a dozen samples was typical for a scientific paper in geochemistry. Today there are likely to be 50 elements on several hundred samples, and the data quantity is so large that it cannot even be published easily by conventional paper publication. Subsequent investigators often do not even have access to the complete data sets, which consisted in part of unpublished data. Individual compilations of published data take large amounts of time to create and are inevitably incomplete, leading to selective and often nonrepresentative comparisons and conclusions. The rise of interdisci-

plinary science also means that scientists in diverse fields need to be able to access data that can be evaluated for completeness and reliability. In addition, it is now recognized that the geochemical data themselves rapidly become outdated unless they are supported by the necessary metadata: exact sample names and locations, analytical methods and errors, archival data, etc. The need for a database that contains all the geochemical and petrological data and supporting metadata in a form in which it can be accessed readily by the scientific community is evident.

[3] We have designed a relational database structure for chemical and petrological data of all types of rock samples that includes all essential metadata. Two major databases are currently using this structure, the RidgePetDB, a petrological database for the ocean floor compiled at the Lamont-Doherty Earth Observatory [*Lehnert et al.*, 1999] and GEOROC, a geochemical database for ocean islands and other oceanic as well as continental rocks compiled at the Max-Planck-Institut für Chemie in Mainz, Germany [*Sarbas et al.*, 1999].

2. Contents of the Database

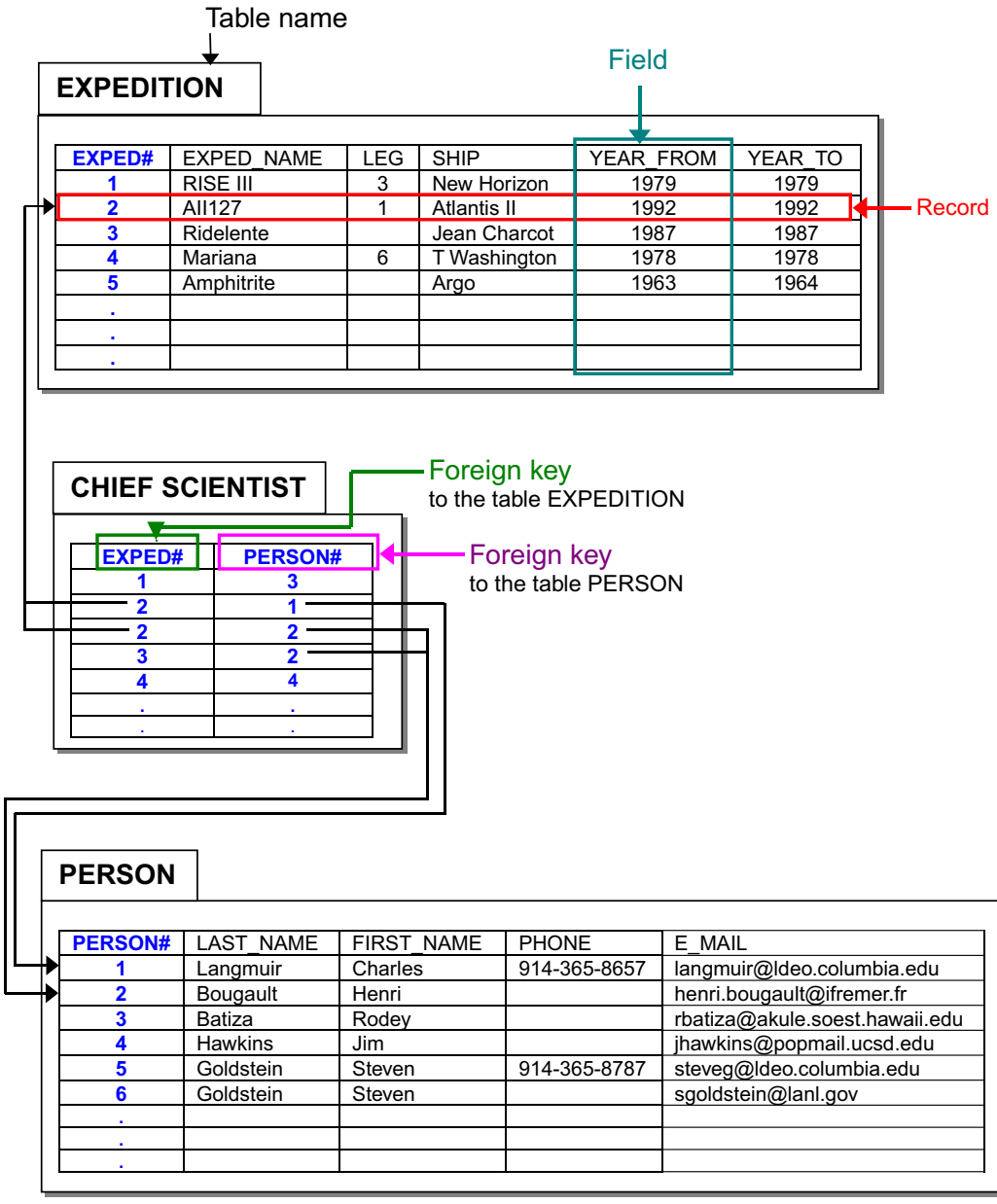
[4] A comprehensive geochemical database has to accommodate far more data than the actual chemical values. Supplementary information describing the data quality as well as the analyzed samples needs to be incorporated because it is essential for proper evaluation of the analytical data and for efficient recovery and sorting of data. Two types of data in the database, therefore, can be distinguished:

1. Primary data comprise all analytical values for rock samples including major oxide and trace element compositions, radiogenic and stable isotope ratios, noble gas contents, and uranium series for different materials like whole rocks, volcanic glasses, mineral phases, and melt or mineral inclusions.
2. Secondary data include (1) analytical metadata that describe the analytical method (technique, laboratory, procedures, errors, precision, standard values, correction procedures) and the material of the sample that was analyzed (glass, whole rock, mineral, etc.), (2) sample metadata that describe the rock sample and its provenance (petrography, age, sampling technique, tectonic setting and geographical position of the sampling site, sampling cruise, navigation method, repository, etc.) and (3) reference metadata that give bibliographical information for the reference which reports the analytical values.

[5] All of the information listed above represents multidimensional, related data that cannot be handled reasonably or efficiently in a two-dimensional flat file format database consisting of a number of separate spreadsheets. Searches for specific data would be complicated and inefficient because data need to be retrieved from each individual file separately. The optimal method for organization and delivery of such complex but related data is a relational database.

3. The Relational Database Model

[6] Figure 1 illustrates the basic concept of relational databases. Data are stored and presented in the form of tables which represent the building blocks of the database. Each table contains a group of logically related data and is identified by a unique name. For example, the table EXPEDITION contains all data relevant to describe expeditions, the table PERSON holds information about persons. A table consists of fields (the vertical columns) and records (the horizontal rows). Each field holds a particular type of information labeled by the



EXPED#	EXPED_NAME	LEG	SHIP	YEAR_FROM	YEAR_TO
1	RISE III	3	New Horizon	1979	1979
2	AII127	1	Atlantis II	1992	1992
3	Ridelente		Jean Charcot	1987	1987
4	Mariana	6	T Washington	1978	1978
5	Amphitrite		Argo	1963	1964
.					
.					
.					

EXPED#	PERSON#
1	3
2	1
2	2
3	2
4	4
.	.
.	.

PERSON#	LAST_NAME	FIRST_NAME	PHONE	E_MAIL
1	Langmuir	Charles	914-365-8657	langmuir@ldeo.columbia.edu
2	Bougault	Henri		henri.bougault@ifremer.fr
3	Batiza	Rodey		rbatiza@akule.soest.hawaii.edu
4	Hawkins	Jim		jhawkins@popmail.ucsd.edu
5	Goldstein	Steven	914-365-8787	steveg@ldeo.columbia.edu
6	Goldstein	Steven		sgoldstein@lanl.gov
.				
.				
.				

Figure 1. Data storage in a relational database. The tables EXPEDITION, CHIEF_SCIENTIST, and PERSON are shown to illustrate the basic units and concepts of a relational database. Data are stored in tables that consist of fields and records. Each table has a primary key which allows unique identification of each record of the table. The primary keys are EXPED# in the table EXPEDITION, PERSON# in the table PERSON, and EXPED# + PERSON# in the table CHIEF_SCIENTIST. Relations among the tables are established through foreign keys. Referencing of primary keys by foreign keys is depicted by arrows.

field name. Each record of a table represents one set of data.

[7] In order to find information in a specific table, it is necessary that each record in a table can be uniquely identified. The unique identifier of a record, called the primary key, is a field or a combination of fields of a table that has a different value in each record. Names or numbers can be used as primary keys. Number keys are generally preferable because queries on numbers work faster and names are often not unique. For example, unique identification of a person in the table PERSON requires a primary key consisting of the fields last name + first name + e-mail or phone number because different persons can have the same first name and last name (see Figure 1). A unique PERSON_NUM is used as the primary key instead.

[8] The various tables of a relational database are related to one another through foreign keys. These are fields or groups of fields that reference the primary key of another table. In Figure 1 the PERSON_NUM in the table CHIEF_SCIENTIST is a foreign key to the table PERSON. It relates an expedition (identified by the foreign key EXPEDITION_NUM) to a person whose name, phone number, and e-mail address are listed in the table PERSON. While primary keys have to be unique by definition, foreign keys usually are not. The same PERSON_NUM will appear more than once in the table CHIEF_SCIENTIST if the person is chief scientist on several expeditions.

[9] The example in Figure 1 reveals some fundamental advantages of the relational model that are crucial for its application to geochemical and petrological data:

1. Redundancy of data is avoided. All data describing a person (first name, last name, etc.) are stored in one record in the table PERSON and are not repeated each time this person is listed in the table CHIEF_SCIENTIST. This principle is applied in general to the geochemical database. For example, method information is entered only once in the table METHOD, rather than a new entry being repeated for every analysis for which the method was used. Location information is entered once for a dredge and is not repeated for every sample from the dredge.
2. Relational databases provide a solution to store multiple values for the same data field. For example, entering two chief scientists for one expedition into a normal spreadsheet would require two columns for CHIEF_SCIENTIST_1 and CHIEF_SCIENTIST_2. If an expedition has a third chief scientist, the spreadsheet would need to be changed by adding a new column CHIEF_SCIENTIST_3 which would propagate to all records in a flat file database. In the relational database, however, an infinite number of chief scientists can be entered for any expedition in the table CHIEF_SCIENTIST. In the same way, it is possible to store a variable number of standard measurements for one method or a variable number of geographical names for one location.

Figure 2. Schema of the geochemical database. For detailed information on table and field content of a particular table and definition of terms used, see http://g-cubed.org. Tables have been grouped according to the type of information that they contain: (a, blue) Tables with sample-related information, (b, yellow) tables with analysis-related information, and (c, orange) tables with reference information. (The tables PERSON and INSTITUTION are also used for sample-related metadata (chief scientists of expeditions, institutions that run expeditions and repositories).) Table names are in white letters on black background, primary keys are in black letters on grey background, and foreign keys are in italics. Arrows point from primary keys to foreign keys depicting relations among the tables.

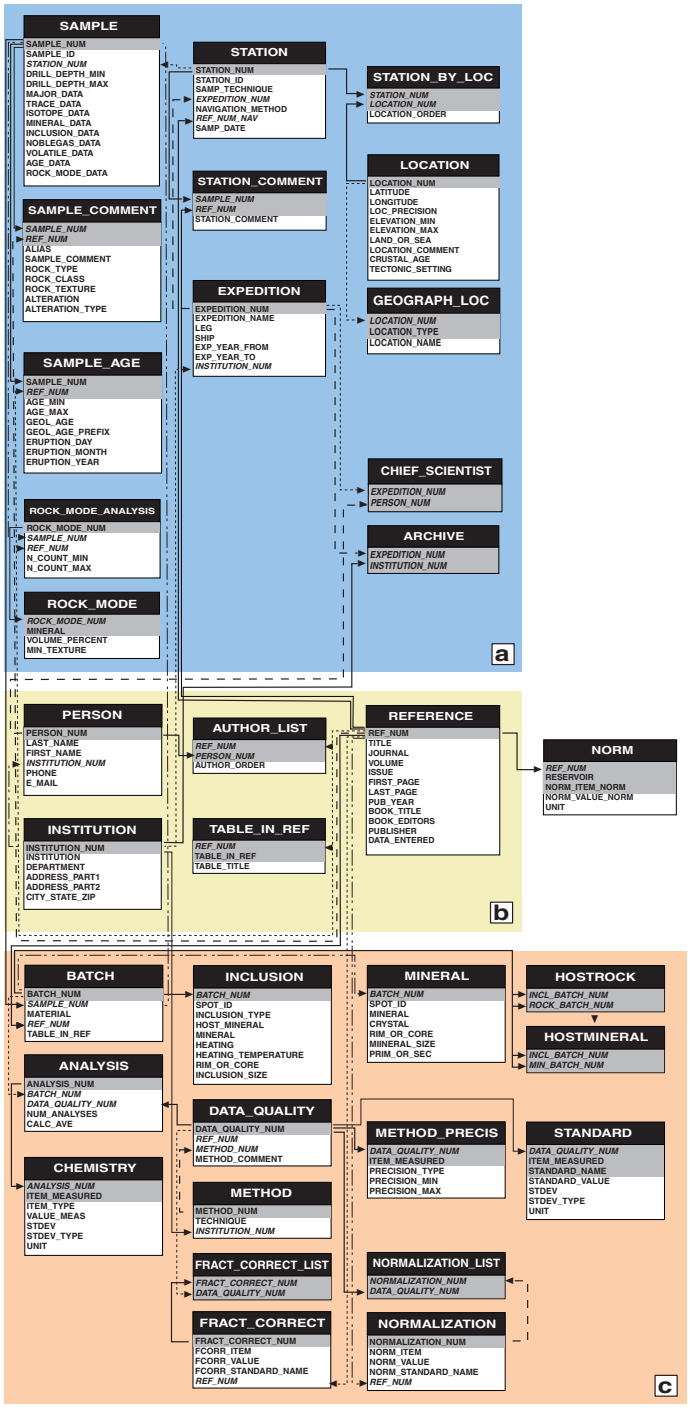

SAMPLE
SAMPLE_NUM
SAMPLE_ID
STATION_NUM
DRILL_DEPTH_MIN
DRILL_DEPTH_MAX
MAJOR_DATA
TRACE_DATA
ISOTOPE_DATA
MINERAL_DATA
INCLUSION_DATA
NOBLEGAS_DATA
VOLATILE_DATA
AGE_DATA
ROCK_MODE_DATA
STATION
STATION_NUM
STATION_ID
SAMP_TECHNIQUE
EXPEDITION_NUM
NAVIGATION_METHOD
REF_NUM_NAV
SAMP_DATE
STATION_BY_LOC
STATION_NUM
LOCATION_NUM
LOCATION_ORDER
LOCATION
LOCATION_NUM
LATITUDE
LONGITUDE
LOC_PRECISION
ELEVATION_MIN
ELEVATION_MAX
LAND_OR_SEA
LOCATION_COMMENT
CRUSTAL_AGE
TECTONIC_SETTING
STATION_COMMENT
SAMPLE_NUM
REF_NUM
STATION_COMMENT
SAMPLE_COMMENT
SAMPLE_NUM
REF_NUM
ALIAS
SAMPLE_COMMENT
ROCK_TYPE
ROCK_CLASS
ROCK_TEXTURE
ALTERATION
ALTERATION_TYPE
EXPEDITION
EXPEDITION_NUM
EXPEDITION_NAME
LEG
SHIP
EXP_YEAR_FROM
EXP_YEAR_TO
INSTITUTION_NUM
GEOGRAPH_LOC
LOCATION_NUM
LOCATION_TYPE
LOCATION_NAME
SAMPLE_AGE
SAMPLE_NUM
REF_NUM
AGE_MIN
AGE_MAX
GEOL_AGE
GEOL_AGE_PREFIX
ERUPTION_DAY
ERUPTION_MONTH
ERUPTION_YEAR
CHIEF_SCIENTIST
EXPEDITION_NUM
PERSON_NUM
ROCK_MODE_ANALYSIS
ROCK_MODE_NUM
SAMPLE_NUM
REF_NUM
N_COUNT_MIN
N_COUNT_MAX
ARCHIVE
EXPEDITION_NUM
INSTITUTION_NUM
ROCK_MODE
ROCK_MODE_NUM
MINERAL
VOLUME_PERCENT
MIN_TEXTURE
a
PERSON
PERSON_NUM
LAST_NAME
FIRST_NAME
INSTITUTION_NUM
PHONE
E_MAIL
AUTHOR_LIST
REF_NUM
PERSON_NUM
AUTHOR_ORDER
REFERENCE
REF_NUM
TITLE
JOURNAL
VOLUME
ISSUE
FIRST_PAGE
LAST_PAGE
PUB_YEAR
BOOK_TITLE
BOOK_EDITORS
PUBLISHER
DATA_ENTERED
NORM
REF_NUM
RESERVOIR
NORM_ITEM_NORM
NORM_VALUE_NORM
UNIT
INSTITUTION
INSTITUTION_NUM
INSTITUTION
DEPARTMENT
ADDRESS_PART1
ADDRESS_PART2
CITY_STATE_ZIP
TABLE_IN_REF
REF_NUM
TABLE_IN_REF
TABLE_TITLE
b
BATCH
BATCH_NUM
SAMPLE_NUM
MATERIAL
REF_NUM
TABLE_IN_REF
INCLUSION
BATCH_NUM
SPOT_ID
INCLUSION_TYPE
HOST_MINERAL
MINERAL
HEATING
HEATING_TEMPERATURE
RIM_OR_CORE
INCLUSION_SIZE
MINERAL
BATCH_NUM
SPOT_ID
MINERAL
CRYSTAL
RIM_OR_CORE
MIINERAL_SIZE
PRIM_OR_SEC
HOSTROCK
INCL_BATCH_NUM
ROCK_BATCH_NUM
HOSTMINERAL
INCL_BATCH_NUM
MIN_BATCH_NUM
ANALYSIS
ANALYSIS_NUM
BATCH_NUM
DATA_QUALITY_NUM
NUM_ANALYSES
CALC_AVE
DATA_QUALITY
DATA_QUALITY_NUM
REF_NUM
METHOD_NUM
METHOD_COMMENT
METHOD_PRECIS
DATA_QUALITY_NUM
ITEM_MEASURED
PRECISION_TYPE
PRECISION_MIN
PRECISION_MAX
STANDARD
DATA_QUALITY_NUM
ITEM_MEASURED
STANDARD_NAME
STANDARD_VALUE
STDEV
STDEV_TYPE
UNIT
CHEMISTRY
ANALYSIS_NUM
ITEM_MEASURED
ITEM_TYPE
VALUE_MEAS
STDEV
STDEV_TYPE
UNIT
METHOD
METHOD_NUM
TECHNIQUE
INSTITUTION_NUM
FRACT_CORRECT_LIST
FRACT_CORRECT_NUM
DATA_QUALITY_NUM
NORMALIZATION_LIST
NORMALIZATION_NUM
DATA_QUALITY_NUM
FRACT_CORRECT
FRACT_CORRECT_NUM
FCORR_ITEM
FCORR_VALUE
FCORR_STANDARD_NAME
REF_NUM
NORMALIZATION
NORMALIZATION_NUM
NORM_ITEM
NORM_VALUE
NORM_STANDARD_NAME
REF_NUM
c

Table 1. Description of Table Contents

Table Name	Table Content
ANALYSES	Lists the chemical analyses in the database. An analysis is defined as a group of items (elements, oxides, isotope ratios) that is reported in one batch (see table BATCHES) and was analyzed with the same procedure (one DATA_QUALITY_NUM).
ARCHIVE	Lists institutions where samples from expeditions are archived by linking the number of an institution to an expedition number. One expedition can have several archives.
AUTHOR_LIST	Lists the authors of a reference by linking a person number to a reference number. There may be several records for a reference if there are several authors.
BATCH	Lists analytical batches in the database. A batch is defined as a group of items (elements, oxides, or isotope ratios) analyzed on the same sample and reported in the same column of a table in a publication. The items can be analyzed by different methods.
CHEMISTRY	Lists the compositional values with standard deviation and unit. There is one record for each item measured (element, oxide, isotope ratio) from an analysis.
CHIEF_SCIENTIST	Lists the chief scientists of expeditions by linking a person number to an expedition number. One expedition can have several chief scientists.
DATA_QUALITY	Describes the data quality of the analyses. The data quality is defined as a method (a technique at a particular laboratory, listed in the table METHOD and referred to by the METHOD_NUM) with a specific procedure (METHOD_COMMENT) used for data in a specific reference.
EXPEDITIONS	Lists all expeditions during which samples were collected. An expedition is uniquely identified by the expedition name plus leg number and has a unique EXPEDITION_NUM.
FRACT_CORRECT	Lists isotope ratios (natural or standard values) that are used for fractionation correction of isotope ratios.
FRACT_CORRECT_LIST	Links the values of an isotope ratio that is used for fractionation correction of a measured isotope ratio to a data quality record.
GEOGRAPH_LOC	Lists geographical names that describe a location. One location can be described by several location types. LOCATION_TYPE is OCEAN, SPREADING_CENTER, ISLAND_GROUP, ISLAND_NAME, LAVA_FLOW, OUTCROP, etc.
HOSTMINERAL	Links analytical data (a batch) of an inclusion to analytical data (a batch) of its host mineral.
HOSTROCK	Links analytical data (a batch) of an inclusion to analytical data (a batch) of its host rock.
INCLUSION	Lists information for batches of inclusions. Information includes type of inclusion (melt or mineral), name of host mineral, description of inclusion's spot that was analyzed, and heating procedures.
INSTITUTION	Lists institutions such as universities, or departments of institutions, with their addresses. Each institution/department is identified by an INSTITUTION_NUM.
LOCATION	Lists locations. A location is any place on Earth defined by its geographical coordinates and its elevation (positive for above sea level or negative for below sea level).
METHOD	Lists methods. A method is defined as a technique, e.g., XRF, EMP, or ICP-MS, performed at a specific institution (laboratory), e.g., LDEO, MIT.
METHOD_PRECIS	Lists information about the precision of analytical data for each item measured with a specific data quality as defined in the table DATA_QUALITY.
MINERAL	Lists information for batches of minerals. Information includes name of mineral, type of grain (phenocryst, etc.), and location of spot that was analyzed.
NORM	Lists values given for the calculated composition of a particular geochemical reservoir in a specific publication that are used for normalization of analytical values.
NORMALIZATION	Lists compositional values of standards that are used for normalization of measured values.
NORMALIZATION_LIST	Links the value of a standard that is used for normalization of analytical values to a data quality record.

Table 1. (continued)

Table Name	Table Content
PERSON	Lists persons that are referred to as authors or chief scientists with their affiliation and contact (phone, e-mail). Each person is uniquely identified by a PERSON_NUM.
REFERENCE	Lists bibliographical information such as title, journal, page numbers, and publication year for references. Each reference is identified by a REF_NUM.
ROCK_MODE	Lists volume percentages of the minerals in a rock mode analyses. There is one row for each mineral of a rock mode analyses.
ROCK_MODE_ANALYSES	Lists modal analyses for samples. A rock mode analyses is defined by the sample number and the number of the reference and is given a unique ROCK_MODE_ANALYSES_NUM. There can be several rock mode analyses for one sample if they were published in different publications.
SAMPLE	Lists all samples in the database for which chemical data are stored. A sample is defined as one piece of rock from a sampling site. Each sample is uniquely identified by the SAMPLE_ID (a database internal code) and has a unique SAMPLE_NUM.
SAMPLE_AGE	Lists ages or age ranges of samples either as absolute ages or by the name of the geological era and also gives dates of eruption for samples from historic lava flows.
SAMPLE_COMMENT	Lists petrographic descriptions and classifications of rock samples by reference. One sample will have several records if described in various publications.
STANDARD	Lists standard values for items measured with the same data quality as analyses from samples that have the same DATA_QUALITY_NUM.
STATION	Lists sampling sites for all samples in the database. A sampling site is defined as a place, an area, or a track where a batch of samples has been collected on a particular expedition that has the same site identification. Examples for sampling sites are a dredge or a DSDP hole. Each sampling site is uniquely described by the STATION_ID and has a unique STATION_NUM.
STATION_BY_LOC	Links sampling sites to locations. A station can have several locations, e.g., an "on" and an "off" location of a dredge or points that define the track of a ship.
STATION_COMMENT	Lists information about sampling sites by reference. One sampling site will have several records if described in various publications.
TABLE_IN_REF	Lists titles of tables as given in a reference.

[10] Development of a relational database structure requires that all data that will be stored in the database are broken down into small logical units corresponding to the tables of the database. Primary keys need to be defined for each table, and relations among the tables need to be established through foreign keys. The logical structure of a relational database is represented in the schema that illustrates tables, fields, and their relations.

4. Structure of Geochemical Database

[11] The schema of the geochemical database is shown in Figure 2. Explanations of table contents and field contents as well as definitions of terms used are available at http://www.g-cubed.org Alphabetical listings of table and field contents are given in Tables 1 and 2, respectively.

4.1. Sample-Related Information

[12] Logically, the central unit of the schema is the table SAMPLE. Each sample for which chemical data are stored in the database is identified by a unique SAMPLE_NUM, which serves as a foreign key throughout the database and links sample metadata in tables such as SAMPLE_COMMENT, STATION, or SAMPLE_AGE to analytical data.

Table 2. Description of Field Contents

Field Name	Field Comment	Example	Format*
ADDRESS_PART1	Street address of the institution	"20 Park Avenue"	CHAR
ADDRESS_PART2	Street address of the institution	"Geoscience Building"	CHAR
AGE_DATA	Shows if age analyses for a sample exist	"Y", "N"	CHAR
AGE_MAX	Maximum age for a sample (in million years)	"0.5"	NUM
AGE_MIN	Minimum age for a sample (in million years)	"0.03"	NUM
ALIAS	Name of the sample in the reference	"D22-1"	CHAR
ALTERATION	Alteration grade of the sample	''F'' = fresh, ''S'' = slightly altered, etc.	CHAR
ALTERATION_TYPE	Type of alteration of the sample	"Serpentinization"	CHAR
ANALYSES_NUM	Number that uniquely identifies an analysis		NUM
AUTHOR_ORDER	Number to identify the position of the person in the list of authors (first author = "1", etc.)		NUM
BATCH_NUM	Number that uniquely identifies a batch		NUM
BOOK_EDITORS	Names of a book's editors	"SAUNDERS-A-D, NORRY-M-J"	CHAR
BOOK_TITLE	Title of a book in which a reference is published	"Magmatism in the Ocean Basins"	CHAR
CALC_AVE	Describes if an analysis is an individual analysis, part of a group of analyses that can be averaged, or an average	"N", "Y", "A"	CHAR
CITY_STATE_ZIP	City, state, and zip code for an institution	"Palisades, NY 10964"	CHAR
COUNTRY	Country of an institution	"USA"	CHAR
CRYSTAL	Type of grain	"PC" = phenocryst, "ML" = microlite, etc.	CHAR
DATA_ENTERED	Indicates if the data of the reference have been entered into the database	"Y", "N"	CHAR
DATA_QUALITY_NUM	Number that uniquely identifies a data quality description		NUM
DEPARTMENT	Name of a department in an address	''Geoscience Department"	CHAR
DRILL_DEPTH_MAX	Maximum drill depth for a sample from a drill core		NUM
DRILL_DEPTH_MIN	Minimum drill depth for a sample from a drill core		NUM
ELEVATION_MAX	Maximum elevation of a sample location above (positive numbers) or below (negative numbers) sea level in meters	"−3462"	NUM
ELEVATION_MIN	Minimum elevation of a sample location above (positive numbers) or below (negative numbers) sea level in meters	"−3375"	NUM
ERUPTION_DAY	Day of eruption for samples from historic eruptions	"28"	NUM
ERUPTION_MONTH	Month of eruption for samples from historic eruptions	"05"	NUM
ERUPTION_YEAR	Year of eruption for samples from historic eruptions	"1988"	NUM
EXPEDITION_NAME	Name of the expedition	"RISE III"	CHAR
EXPEDITION_NUM	Number that uniquely identifies an expedition defined by the name of the expedition and the leg		NUM
EXP_YEAR_FROM	Year the expedition started	"1986"	NUM
EXP_YEAR_TO	Year the expedition ended	"1987"	NUM
E_MAIL	E-mail address of the person	"lehnert@ldeo.columbia.edu"	CHAR

Table 2. (continued)

Field Name	Field Content	Example	Format
FCORR_ITEM	Name of the isotope ratio that is used for the fractionation correction	"SR88_SR86"	CHAR
FCORR_STANDARD_NAME	Name of the standard whose values have been used for a fractionation correction in case it is not a natural isotope ratio	"NBS981"	CHAR
FCORR_VALUE	Value for the isotope ratio that is used for the fractionation correction		NUM
FIRST_NAME	First name of a person	"John B"	CHAR
FIRST_PAGE	Number of the first page		NUM
FRACT_CORRECT_NUM	Number that uniquely identifies a fractionation correction procedure		NUM
GEOL_AGE	Geological age of a sample	"Tertiary"	CHAR
GEOL_AGE_PREFIX	Prefix for the geological age of a sample	"Upper"	CHAR
HEATING	Describes if a glass inclusion was heated for analysis	"Y", "N"	CHAR
HEATING_TEMPERATURE	Temperature in degrees Celsius if HEATING="Y"	"1020"	CHAR
HOST_MINERAL	Name of host mineral for the inclusion	"PLAG"	CHAR
INCLUSION_DATA	Shows if inclusion analyses for the sample exist	"Y", "N"	CHAR
INCLUSION_SIZE	Size of the melt or mineral inclusion	"10 x 30 um" u = microns	CHAR
INCLUSION_TYPE	Type of inclusion	"GL"(glass)	CHAR
INCL_BATCH_NUM	Number that identifies the batch of an inclusion, foreign key to the table BATCH		NUM
INSTITUTION	Name of the institution in an address	"University of Rhode Island"	CHAR
INSTITUTION_NUM	Number that uniquely identifies the address of a department/institution		NUM
ISOTOPE_DATA	Shows if isotope data for the sample exists	"Y", "N"	CHAR
ISSUE	Number of the journal's issue	"12B"	CHAR
ITEM_MEASURED	Name of element, oxide, or isotope ratio	"Cr", "SiO2", "Pb206_Pb204"	CHAR
ITEM_TYPE	Group name for elements/oxides/isotope ratios	"REE", "MAJ"	CHAR
JOURNAL	Name of the journal for the reference	"Earth Planet Sci Lett"	CHAR
LAND_OR_SEA	Describes if a sample is submarine or subaerial	"SAQ" (subaquatic), "SAE" (subaerial)	CHAR
LAST_NAME	Last name of a person	"Langmuir"	CHAR
LAST_PAGE	Number of the last page		NUM
LATITUDE	Latitude value for a location in decimal degrees, negative values for south latitudes	"35.145"	NUM
LEG	Number of the leg		NUM
LOCATION_COMMENT	Remarks on a location	"Inner wall of axial rift"	CHAR
LOCATION_NAME	Name of geographical/tectonic feature	"East Pacific Rise", "Lamont Seamount", "Pacific"	CHAR
LOCATION_NUM	Number that uniquely identifies a geographic location defined by latitude/longitude/elevation		NUM
LOCATION_ORDER	Number that shows in which order locations have been visited if the sampling site is a track; e.g., the on site of a dredge = 1, the off site = 2		NUM
LOCATION_TYPE	Type of geographical/tectonic feature that is described by the LOCATION_NAME	"Spreading Center", "Seamount", "Ocean"	CHAR
LOC_PRECISION	Number that gives the precision of the latitude and longitude values	"0.001"	NUM

Table 2. (continued)

Field Name	Field Content	Example	Format
LONGITUDE	Longitude value in decimal degrees, negative values for west longitudes	"−109.073"	NUM
MAJOR_DATA	Shows if major element data for the sample exist	"Y", "N"	CHAR
MATERIAL	Material that has been analyzed (whole rock, glass, mineral, or inclusion)	"WR", "GL", "MIN", "INCL"	CHAR
METHOD_COMMENT	Remarks on the data quality, e.g., description of a special procedure	"Leached in 6N HCl"	CHAR
METHOD_NUM	Number that uniquely identifies a method		NUM
MINERAL	Name of a mineral	"PLAG", "OL", etc.	CHAR
MINERAL_DATA	Shows if mineral analyses for the sample exist	"Y", "N"	CHAR
MINERAL_SIZE	Size of the mineral grain	"1 mm diameter"	CHAR
MIN_BATCH_NUM	Number that identifies the batch of the host mineral, foreign key to the table BATCH		NUM
MIN_TEXTURE	Textural remarks about the mineral	"Euhedral"	CHAR
NAVIGATION_METHOD	Method of navigation used to determine the location of a station on an expedition	"SAT", "GPS", "SB" (SEABEAM), etc.	CHAR
NOBLEGAS_DATA	Shows if noble gas analyses for a sample exist	"Y", "N"	CHAR
NORMALIZATION_NUM	Number that uniquely identifies a normalization		NUM
NORM_ITEM	Name of item (element/oxide/isotope ratio) that is normalized	"SiO2"	CHAR
NORM_ITEM_NORM	Name of item (element/oxide/isotope ratio) that is normalized in the table NORM	"Nb", "MgO"	CHAR
NORM_STANDARD_NAME	Name of the standard	"JDF-D2"	CHAR
NORM_VALUE	Value of the standard for the item measured that is normalized		NUM
NORM_VALUE_NORM	Value of the reservoir in the table NORM that is used for normalization		NUM
NUM_ANALYSES	Number of analyses that have been averaged (when CALC_AVE = "A")		NUM
N_COUNT_MAX	Maximum number of points counted for a rock mode analysis		NUM
N_COUNT_MIN	Minimum number of points counted for a rock mode analysis		NUM
PERSON_NUM	Number that uniquely identifies a person		NUM
PHONE	Phone number of the person	"914-359-5137"	CHAR
PRECISION_MAX	Maximum value for the precision		NUM
PRECISION_MIN	Minimum value for the precision		NUM
PRECISION_TYPE	Explains how the precision for the data quality is described	"2SREL" = 2 sigma relative standard deviation	CHAR
PRIM_OR_SEC	Describes if the mineral is primary or secondary	"P", "S"	CHAR
PUBLISHER	Name of the book's publisher	"ELSEVIER"	CHAR
PUB_YEAR	Year of publication of the reference	"1979"	NUM
REF_NUM	Number that uniquely identifies a reference		NUM
RESERVOIR	Acronym or name for a geochemical reservoir	"EM1"	CHAR
RIM_OR_CORE	Describes if the analytical spot was at the rim or the core of the inclusion	"C", "R", "I"	CHAR
ROCKMODE_DATA	Shows if a modal analysis for the sample exists	"Y", "N"	CHAR
ROCK_BATCH_NUM	Number that identifies the batch of the host rock, foreign key to the table BATCH		NUM
ROCK_CLASS	Rock classification of the sample	"Basalt", "Gabbro"	CHAR
ROCK_MODE_NUM	Number that uniquely identifies a rock mode analysis		NUM
ROCK_TEXTURE	Petrographic description of the sample	"Aphyric, vesicular"	CHAR

Table 2. (continued)

Field Name	Field Content	Example	Format
ROCK_TYPE	Describes if the sample is volcanic, subvolcanic, plutonic, metamorphic, or a mantle rock	"V" = volcanic, "I" = plutonic, etc.	CHAR
SAMPLE_COMMENT	Remarks on the sample	"Pillow basalt with glass rim"	CHAR
SAMPLE_ID	Database name given to the sample to allow its unique identification based on the expedition during which the sample was taken, number of the sampling site (e.g., a dredge station), and the sample number.	"TRI0101-036-009" (dredge 36, sample 9 from the Trident expedition 101)	CHAR
SAMPLE_NUM	Number that uniquely identifies a sample		NUM
SAMP_DATE	Date on which a sampling site was visited	"12-01-87"	CHAR
SAMP_TECHNIQUE	Technique by which a sample has been collected	"DR", "SUB" (Submersible), "OC" (Outcrop), etc.	CHAR
SHIP	Name of the ship	"Robert Conrad"	CHAR
SPOT_ID	Identification of an analytical spot for mineral and inclusion analyses	"2B"	CHAR
STANDARD_NAME	Name of the standard	"BCR-1"	CHAR
STANDARD_VALUE	Value analyzed for the standard		NUM
STATION_COMMENT	Remarks on a sampling site, e.g., weight and content of a dredge	"40 kg of fresh pillow, some boulders of lherzolite"	CHAR
STATION_ID	Database name given to a sampling site to allow its unique identification. For undersea samples this is based on the expedition during which the site was visited and its number (e.g., a dredge station number)	"ARGAMPH-006" (dredge 6 from the cruise Amphitrite on the R/V Argo)	CHAR
STATION_NUM	Number that uniquely identifies a sampling site		NUM
STDEV	Absolute standard deviation given for averaged values		NUM
STDEV_TYPE	Describes how the standard deviation of a compositional value is reported	"REL", "ABS"	CHAR
TABLE_IN_REF	Number of the table in the publication in which the batch is listed		NUM
TABLE_TITLE	Table caption	"Major element data for rocks from the EPR 9 deg N"	CHAR
TECHNIQUE	Analytical technique	"EMP", "XRF", "ICP-MS", etc.	CHAR
TITLE	Title of a reference	"Magmatic Evolution of the Easter Microplate"	CHAR
TRACE_DATA	Shows if trace element data for the sample exist	"Y", "N"	CHAR
UNIT	Unit of a compositional value	"WT%", "PPM"	CHAR
VALUE_MEAS	Compositional value		NUM
VOLATILE_DATA	Shows if volatile analyses for the sample exist	"Y", "N"	CHAR
VOLUME	Number of the journal's volume	"129A"	CHAR
VOLUME_PERCENT	Volume percentage of the mineral in the rock mode		NUM

*CHAR, character; NUM, number.

Table 3. Verbal Descriptions of Locations

LOCATION_TYPE	LOCATION_NAME
OCEAN	PACIFIC
ISLAND_GROUP	HAWAIIAN ISLANDS
OCEAN_ISLAND	HAWAII
VOLCANO	KILAUEA
LAVA_FLOW	PUNA BASALT
OUTCROP	OLD VISITOR CENTER

[13] Each sample in the table SAMPLE is linked to metadata regarding sampling method and geographical location via the foreign key STATION_NUM. A STATION is defined as a sampling site (e.g., a dredge, a drill hole, or an outcrop) and is described by parameters such as sampling technique, navigation method, date of sampling, and expedition. A LOCATION is a static point on the Earth's globe defined by latitude, longitude, and elevation. Stations and locations are linked in the table STATION_BY_LOCATION. This table allows a station to have more than one location, for example, an on and an off site for a dredge or a sequence of locations for a ship track, and a location to have more than one station in case the same sampling site has been visited by several expeditions, e.g., a Deep Sea Drilling Project (DSDP) hole that has been revisited.

[14] Verbal descriptions of locations, that is, geographical names ranging from the name of an ocean to the name of an outcrop, are stored in the table GEOGRAPH_LOC. Geographical or tectonic features such as "ocean," "plate," "spreading center," or "ocean island" are variables entered in the field LOCATION_TYPE and have a LOCATION_NAME. In this way any geographical feature, whether it is on continents or on the ocean floor, can be included, and as many location types as desired can be used to describe a location. See Table 3.

4.2. Problem of Unique Sample Identification

[15] Unique identification of samples is essential for the usefulness of the database. Valuable information is lost if different data sets for the same sample cannot be combined because the sample cannot be unambiguously recognized in the references.

[16] Sample names from publications are often not unique. For example, a name like DR22-1 (rock fragment 1 from a dredge 22) has been given to samples collected on different cruises. On the other hand, the sample DR22-1 from a particular cruise (e.g., the PASCUA cruise of the R/V *Washington*) can have different names in different references, e.g., PAS22-1, P22-1, or WAS22/1, or even a personal name assigned to it that bears no relation to its origin. The development and use of unique sample identifiers is a clear need for all future sample collection and publication.

[17] We have included the field SAMPLE_ID in the table SAMPLE, which should contain a unique and meaningful name for each sample. Unique names can be created for samples based on parameters such as the time they were sampled, the person who sampled them (or the name of an expedition or cruise), and the location where they were sampled.

[18] Many ocean floor samples have unique names assigned to them upon collection,

associated with a ship, a year, and a station number. All other samples need to be renamed in order to take advantage of the SAMPLE_ID in the database. This is possible only if the pertinent information such as sampling year or sampling person is accessible. For most submarine samples the information is included in publications or can be obtained from authors or other sources. More than 98% of the samples in the RidgePetDB have a unique SAMPLE_ID, which allows all data for individual samples from different laboratories to be integrated.

[19] In contrast, names of subaerial samples are not systemized, and metadata on sampling expedition or sampling year are usually lacking. Often even the exact location of the sample is obscure, and a unique SAMPLE_ID cannot be created. In the GEOROC database which covers mostly subaerial samples (ocean islands and continents) the SAMPLE_ID could therefore not be applied. Instead, the sample name from the original publication is entered as the SAMPLE_ID, which makes the SAMPLE_ID nonunique. Different data sets for one sample can only be joined by the user by means of the detailed information on the sample's provenance. There is a clear need for a universal sample-naming protocol for subaerial samples.

4.3. Analysis-Related Information

[20] All analytical data are stored in one single table, the table CHEMISTRY. Each row in the table CHEMISTRY contains one analytical value. The name of the oxide, element, or isotope ratio for a given value is a variable entered in the field ITEM_MEASURED.

[21] This design differs from familiar geochemical data tables, where all the data for one sample are in one row or column, and offers major advantages for the functionality and applicability of the database:

1. By making the item measured a variable, any new item can be entered without the need to modify existing tables and completely redo the database structure.
2. Data storage is much more efficient in a table that contains one field for each item measured. In flat presentation a large number of fields would be empty because a record (an analysis) usually includes only data for a few of the items.

[22] Figure 3 illustrates the way analytical values are linked to the sample by means of the BATCH_NUM and the ANALYSES_NUM, and to metadata relevant to the data

Figure 3. Organization of analytical data and metadata in the geochemical database. How information from a data table in a reference is stored in the database is illustrated. Each column in the data table is defined as a BATCH, e.g., the column for sample D1-1 is BATCH# 100. Each batch is linked to a sample by the SAMPLE number, which in turn provides the link to all sample metadata. Chemical data in a batch that were analyzed by the same method are grouped as ANALYSES. For example, all rare earth elements (REE) analyzed by instrumental neutron activation analysis (INAA) at Australian National University (ANU) on sample D1-1 have ANALYSIS# 50, the trace elements Sr, Rb, Sm, Nd that were analyzed by isotope dilution mass spectrometry on sample D1-1, have ANALYSIS# 51, etc. The BATCH# allows the reconstruction of the original data table by linking the various analyses of a sample in a data table. A D_Q# (data quality number) is listed with each analyses to provide the link to information about the data quality. A "data quality" in the table DATA_QUALITY is defined as a method described in the table METHOD (METHOD#) and a METHOD_COMMENT that describes special procedures such as "leached in 6 N HCl." Data in the example are from three different methods (INAA at ANU, isotope dilution mass spectrometry (MS-ID) at Scripps Institution of Oceanography (SIO), and MS at SIO), but five different data quality records are necessary because MS-ID and MS have been performed both on unleached and leached samples.

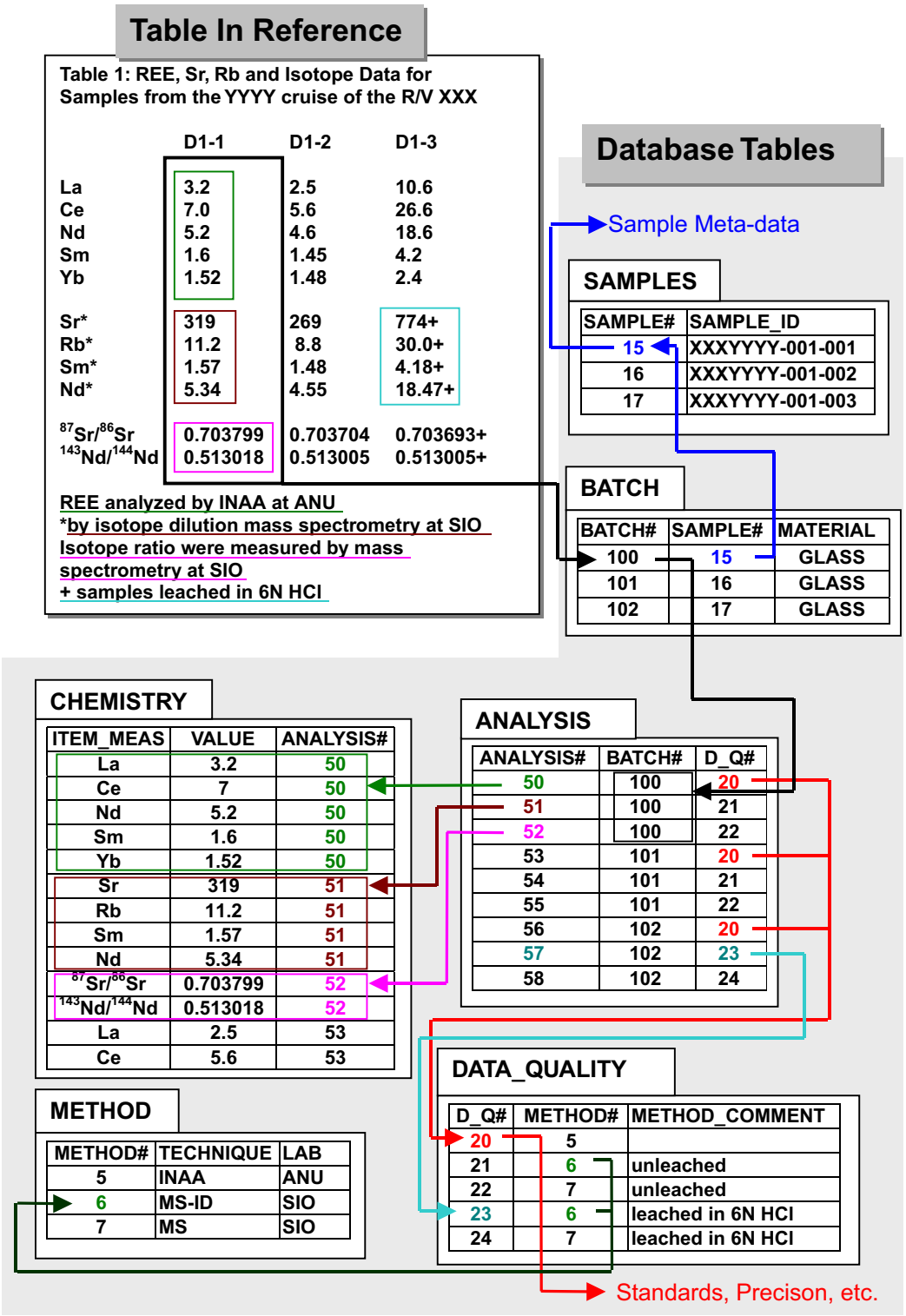


Table In Reference
Table 1: REE, Sr, Rb and Isotope Data for Samples from the YYYY cruise of the R/V XXX
D1-1 D1-2 D1-3
La 3.2 2.5 10.6
Ce 7.0 5.6 26.6
Nd 5.2 4.6 18.6
Sm 1.6 1.45 4.2
Yb 1.52 1.48 2.4
Sr* 319 269 774+
Rb* 11.2 8.8 30.0+
Sm* 1.57 1.48 4.18+
Nd* 5.34 4.55 18.47+
87Sr/86Sr 0.703799 0.703704 0.703693+
143Nd/144Nd 0.513018 0.513005 0.513005+
REE analyzed by INAA at ANU
*by isotope dilution mass spectrometry at SIO
Isotope ratio were measured by mass spectrometry at SIO
+ samples leached in 6N HCl
Database Tables
Sample Meta-data
SAMPLES
SAMPLE# SAMPLE_ID
15 XXXYYYY-001-001
16 XXXYYYY-001-002
17 XXXYYYY-001-003
BATCH
BATCH# SAMPLE# MATERIAL
100 15 GLASS
101 16 GLASS
102 17 GLASS
CHEMISTRY
ITEM_MEAS VALUE ANALYSIS#
La 3.2 50
Ce 7 50
Nd 5.2 50
Sm 1.6 50
Yb 1.52 50
Sr 319 51
Rb 11.2 51
Sm 1.57 51
Nd 5.34 51
87Sr/86Sr 0.703799 52
143Nd/144Nd 0.513018 52
La 2.5 53
Ce 5.6 53
ANALYSIS
ANALYSIS# BATCH# D_Q#
50 100 20
51 100 21
52 100 22
53 101 20
54 101 21
55 101 22
56 102 20
57 102 23
58 102 24
DATA_QUALITY
D_Q# METHOD# METHOD_COMMENT
20 5
21 6 unleached
22 7 unleached
23 6 leached in 6N HCl
24 7 leached in 6N HCl
METHOD
METHOD# TECHNIQUE LAB
5 INAA ANU
6 MS-ID SIO
7 MS SIO
Standards, Precison, etc.

quality by means of the DATA_QUALITY_NUM. Each record in the table DATA_QUALITY describes one method, that is, a specific technique used at a specific laboratory, with a certain procedure, for example, "mass spectrometry at SIO, samples leached in 6 N HCl."

[23] The table BATCH contains the field MATERIAL that describes which material was analyzed. Currently, whole rocks, volcanic glasses, minerals, and inclusions (melt, mineral) are included in the database, but any other material, e.g., fluid inclusions, can be added. Information about the material is given in separate tables. For example, the table MINERAL supplies the name of the mineral, the type of grain that was analyzed (phenocryst, groundmass crystal, xenocryst, etc.), and the location of the analytical spot (rim or core). New tables can easily be added for additional materials.

4.4. Reference-Related Information

[24] All data in the database are linked to a reference listed in the table REFERENCES. Authors are given in the AUTHOR_LIST where a REF_NUM is linked to one or more PERSON_NUM(s). The table PERSON lists persons referenced in the database as authors or chief scientists with contact information and affiliation (INSTITUTION_NUM). The table INSTITUTIONS contains departments of institutions that are referenced as persons' affiliation, which serve as sample repositories in the table ARCHIVE or were organizing institutions of expeditions in the table EXPEDITION.

4.5. Table NORM

[25] The purpose of the table NORM is to store chemical values that can be used to normalize analytical values to the composition of hypothetical chemical reservoirs as given in the literature.

5. Operational Aspects of the Database

[26] The schema is the conceptual structure of a database. In order to build a functional database, the physical structure of the database needs to be created with a relational database management system (RDBMS). The database then needs to be populated with data, and finally interfaces have to be developed to allow users to work with the data.

[27] Any RDBMS can be used to implement the geochemical database. We have successfully implemented the database in ORACLE under UNIX, POSTGRES under LINUX, and MS ACCESS under WINDOWS (NT, 95, 98).

[28] Data entry forms and applications are essential to facilitate the complex process of loading data into the tables while ensuring the uniqueness of primary keys, data integrity, and correct referencing of primary keys by foreign keys. Different types of data entry forms have been developed for the current applications of the geochemical database structure and will be described in subsequent publications.

[29] Access to the database for searching, viewing, and downloading data is best provided over the Internet in order to make it independent of computer platforms and locations. For the current applications, Web interfaces have been developed using Active Server Pages technology. The interfaces allow the user to select samples and data by means of graphical tools such as menus, scrollable lists, or buttons that generate and submit dynamic SQL statements to the database [*Lehnert et al.*, 1999; *Nohl et al.*, 1999; Y. Su et al., PETDB on the World Wide Web: A comprehensive method for accessing petrological data over the Internet, manuscript in preparation, 2000].

6. Conclusions

[30] The database structure presented here is a very flexible entity with a broad range of possible applications to Earth science data. The database has the following general characteristics:

1. Comprehensiveness allows the database to accommodate any type of analytical data for rocks, minerals, and inclusions (and other materials if desired) as well as all significant metadata.
2. All data are accessible. Relations in the database have been established such that the content of any particular field in the database can be used as a criterion to retrieve information from all other fields in all other tables. This ensures that the user can access the entire range of metadata to constrain his selection of samples or data, which makes searches in the database efficient.
3. There is minimum data redundancy. Unnecessary repetition of data has been avoided by extracting information that is not exclusively dependent on the primary key of a table into separate tables (normalization).
4. Data storage is efficient. Tables have been designed in a way that there is a minimum of empty unused fields.
5. Flexibility allows the broad application of the database, which has been assured by keeping the use of tables and fields as general as possible. The database structure could be used readily for further materials and for samples from all tectonic settings.

[31] It is hoped that the design and implementation of such a database for geochemical and petrological data will influence how data are collected and reported in the literature. The database will reveal the lack of unique sample names, complete data sets, and essential metadata in the existing literature, and it may contribute to more effective data collection, presentation, and storage in future publications.

Acknowledgments

[32] The design of the database structure has benefited from discussions with W.F. Ryan, W. Jin, A. Hofmann, P. Friberg, and E. Matthews. We would like to thank all of them for their valuable input. This work was supported at Lamont by a grant from the National Science Foundation.

References

Lehnert, K. L., Y. Su, and C. H. Langmuir, The RIDGE petrological database on the World Wide Web (abstract), *Eos Trans. AGU*, *80*(46), Fall Meet. Suppl., F1183, 1999.

Nohl, U., and B. Sarbas, GEOROC, the MPI geochemical rock database: Introduction of the Web interface (abstract), *Eos Trans. AGU*, *80*(46), Fall Meet. Suppl., F1177, 1999.

Sarbas, B., K. P. Jochum, U. Nohl, and A. W. Hofmann, GEOROC, the MPI Geochemical Rock Database: A new tool for geochemists (abstract), *Eos Trans. AGU*, *80*(46), Fall Meet. Suppl., F1184, 1999.