CONF-9409423--1

AN IMPLEMENTATION OF SAS IN AN ENVIRONMENTAL INFORMATION SYSTEM

Teresa L. James
University of Tennessee, Knoxville, Tennessee 37996

Beverly C. Zygmunt
Oak Ridge National Laboratory*
P.O. Box 2008
Oak Ridge, Tennessee 37831

To be presented at the Southeastern SAS User Group Conference, Charleston, South Carolina, September 18-20, 1994; to be published in the Proceedings.

*Managed by Martin Marietta Energy Systems, Inc., under contract DE-AC05-84OR21400 with the U.S. Department of Energy.

An Implementation of SAS® in an Environmental Information System

Teresa James, University of Tennessee, Knoxville, TN 37996
Beverly Cather Zygmunt, Oak Ridge National Laboratory[1], Oak Ridge, TN 37831

ABSTRACT

This paper describes a software environmental database information system that uses SAS to process data and ORACLE® as the relational database management system (RDBMS). The hardware includes a network of UNIX-based servers and workstations. The relational database consists of large tables containing environmental measurement data, as well as other smaller tables with reference, metadata and internal administrative information. The data come in a variety of formats and must be converted to conform to the system's standards. SAS/ACCESS® and PROC SQL are used extensively in the data processing.

INTRODUCTION

The three U.S. Department of Energy (DOE) installations on the Oak Ridge Reservation (Oak Ridge National Laboratory, Y-12, and K-25) were established during World War II as part of the Manhattan Project that "built the bomb." That research, and work in more recent years, has resulted in the generation of radioactive materials and other toxic wastes. The Oak Ridge Environmental Information System (OREIS) was initiated to provide a consolidated repository of environmental data from the Oak Ridge facilities and DOE plants in Portsmouth, Ohio, and Paducah, Kentucky. (Martin Marietta Energy Systems manages the Oak Ridge installations and the Environmental Restoration programs at the latter two sites.)

The primary use of OREIS data is to provide access to project results by regulators. A secondary use is to serve as background data for other projects. Data within OREIS come from a number of sources, including groundwater, surface water, sediments, soils, air, and biota. Associated with the data types are extensive descriptive and qualifier metadata, whose purpose is to define the quality of the data and thus permit end users to determine the appropriateness of the data for their purposes.

Most OREIS data are generated by projects that deal with environmental compliance, surveillance, and restoration activities. The OREIS user community includes federal and state regulators, data managers and other staff from environmental restoration projects, environmental compliance personnel and monitoring staff, and staff working on biological assessment activities.

One of the major efforts in OREIS is to receive the data from environmental projects and make them available to the users. This paper will briefly describe some key relational database concepts, the relational structure of the OREIS tables, and the processing that takes place to convert the data into the OREIS structure.

THE OREIS OPERATING ENVIRONMENT

The OREIS hardware includes a networked system of Sun servers, Sun and other UNIX-based workstations, X terminals, and PCs. The OREIS RDBMS is ORACLE. For security purposes, ORACLE is installed on a database server that resides behind its own network bridge. Only SQL calls from trusted clients (i.e., OREIS workstations) are allowed to pass through the bridge. PC and X-terminal clients make their SQL calls via a connection to one of the OREIS workstations.

The data are received in a variety of formats, reformatted and reviewed using SAS and SAS/ACCESS, and are then loaded into an ORACLE database. From there, they can be accessed using the integrated OREIS software,

---

[1]Managed by Martin Marietta Energy Systems, Inc., for the U.S. Department of Energy under contract DE-AC05-84OR21400.

which includes ORACLE Browser®, SAS/ASSIST®, SAS/INSIGHT®, and ARC/INFO®.

For security reasons, the ORACLE passwords cannot be embedded within the SAS programs used to process the data. Therefore, the following code is included in these programs and used to prompt for the ORACLE user name and password (not displayed). The ORACLE TWO_TASK variable is obtained for use in the connect string.

```
%let user=;
%let password=;
%window pass color=blue
 #1 @1 "Enter your Username"
attr=highlight color=yellow
 @22 user 4 attr=underline display=yes
required=yes
 #3 @1 "Enter your Password"
attr=highlight color=yellow
 @22 password 7 attr=underline
display=no required=yes;

%display pass;
%let twotask="@%sysget(TWO_TASK)";
```

RDBMS TERMINOLOGY

The following are common RDBMS terms that are included to aid you in understanding the database concepts discussed in the paper.

- **SQL** (pronounced "sequel") stands for structured query language. It is an English-like language used to communicate with ORACLE and other database systems.

- Data in an RDBMS are stored in **tables**. A table consists of **columns** (also known as **fields**) and **rows**. A **record** is a single row of data. An RDBMS table corresponds to a SAS data set; a field to a SAS data set variable; and a row to a SAS data set observation.

- **Integrity constraints** provide a mechanism for defining business rules for a column. An integrity constraint is a statement about a column's data that is always true. An RDBMS will not allow you to load data into a table if an integrity constraint would be violated in the process.

- A **primary key** (PK) integrity constraint will not permit duplicate values or nulls in a column. A **foreign key** (FK) constraint requires that a value in a column match a value in a related table's primary key. The foreign key/primary key association is sometimes called a **parent/child** relationship.

THE DATA MODEL

The OREIS database was modeled using ORACLE's computer aided software engineering (CASE) products. Because of SAS restrictions, each field name in the OREIS tables, even though it may be up to 31-characters long, is unique in the first eight characters. In other words, if an ORACLE table is brought into SAS, the name of each resultant SAS variable will be the first eight characters of the ORACLE field name.

The OREIS tables are divided into five categories: data, reference, metadata, administrative, and change. (See Figure 1.)

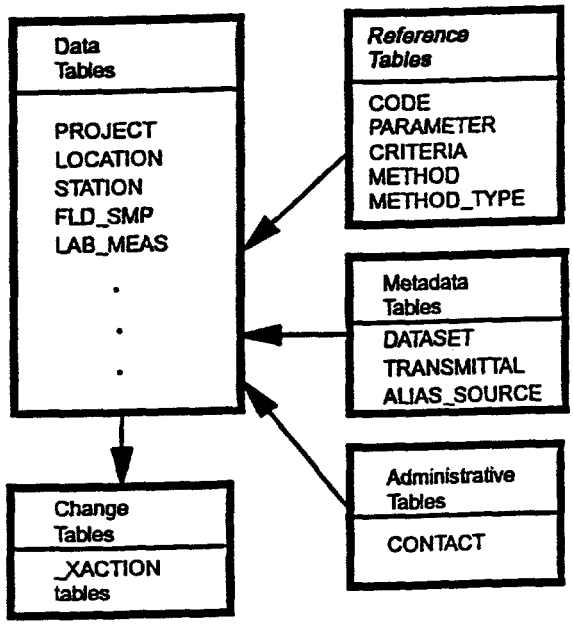

Figure 1

The data tables comprise the largest category and include PROJECT, LOCATION, STATION, WELL, WELLCONS, BORECONS, LITHOLOGY, BIOTA, FLD_EVENT, FLD_MEAS, FLD_SMP, LAB_MEAS, LAB_SMP, QC_FLD_SMP, QC_LAB_MEAS, QC_LAB_SMP, and ASSOCIATE_QC . (See Figure 2.)

The PROJECT table contains records pertaining to the Environmental Restoration (ER) projects. This table contains (as do most other tables) an OREIS-assigned primary key. These keys are discussed in the next section.

Each record in LOCATION refers to a physical point on the earth. The primary key (LOCATION_ID) of this table is the unique identifier used by the OREIS geographic information system (GIS) ARC/INFO. Because each of the physical locations represented by the records in LOCATION may be known by various names to the different projects, there is a one-to-many relationship between the LOCATION and STATION tables. The STATION table includes the name and type of station, as defined by a project.

The WELL table is the "parent" table of both WELLCONS and BORECONS. It contains general construction information. The information it contains applies to both wells and boreholes. The WELLCONS and BORECONS tables contain more detailed construction information. The former also includes additional data pertaining to the development of the well. The LITHOLOGY table is the "child" of BORECONS.

A number of the data tables contain sampling and measurement results, usually with multiple records per sampling station. The primary key of STATION (STATION_ID) exists as a foreign key in BIOTA, FLD_EVENT and FLD_SMP. Because of the parent/child relationships, this same information can be matched to records in FLD_MEAS, LAB_MEAS, and LAB_SMP. The child of FLD_EVENT is FLD_MEAS.

QC samples, such as trip blanks, field blanks, and equipment rinsates, are usually associated with more than one field sample. ASSOCIATE_QC provides a method of joining QC_FLD_SMP and FLD_SMP, so that QC samples can be linked to their corresponding laboratory samples.The QC tables QC_LAB_SMP and QC_LAB_MEAS are virtually identical in structure to the corresponding non-QC tables LAB_SMP and LAB_MEAS.

The reference tables CRITERIA, CODE, PARAMETER, METHOD, and METHOD_TYPE supply additional information that can be linked to the data tables. In some cases, these tables provide checks on the data as they are entered into the OREIS tables. For example, a record cannot be loaded into the LAB_MEAS table if it contains a PARAMTR value not found in PARAMETER. This restriction is a result of the foreign key integrity constraint that has been established between the two tables.

As their name implies, the metadata tables contain information about the data. DATASET includes cautions about using the data, background information on why the data were collected, what processing was done to them, and applications for which they have been used. The TRANSMITTAL table contains information such as medium description, the number of records from the transmittal loaded into the database, and resolution of problems encountered in the processing of the data. ALIAS_SOURCE gives the mapping between source data sets and variables to OREIS tables and fields.

The administrative table is CONTACT. It provides information on all persons associated with OREIS: data generators, data custodians, OREIS staff, and OREIS users.

The OREIS tables contain processed, approved data. Any changes to any of the records in the OREIS tables are documented by using the change (XACTION) tables. The change tables are identical to the OREIS tables, with the addition of fields that define who authorized the change, the type of change (such as update or deletion), and comments associated with the change.

THE TRANS TABLES

TRANS tables are copies of all of the OREIS data tables (except PROJECT and the XACTION tables). The data in the TRANS area have not been QA'd or approved. Because the TRANS area is a "working area," there are no integrity constraints on the TRANS tables except for the primary keys. Data processors add and delete data from these tables freely as needed during the data processing and review.

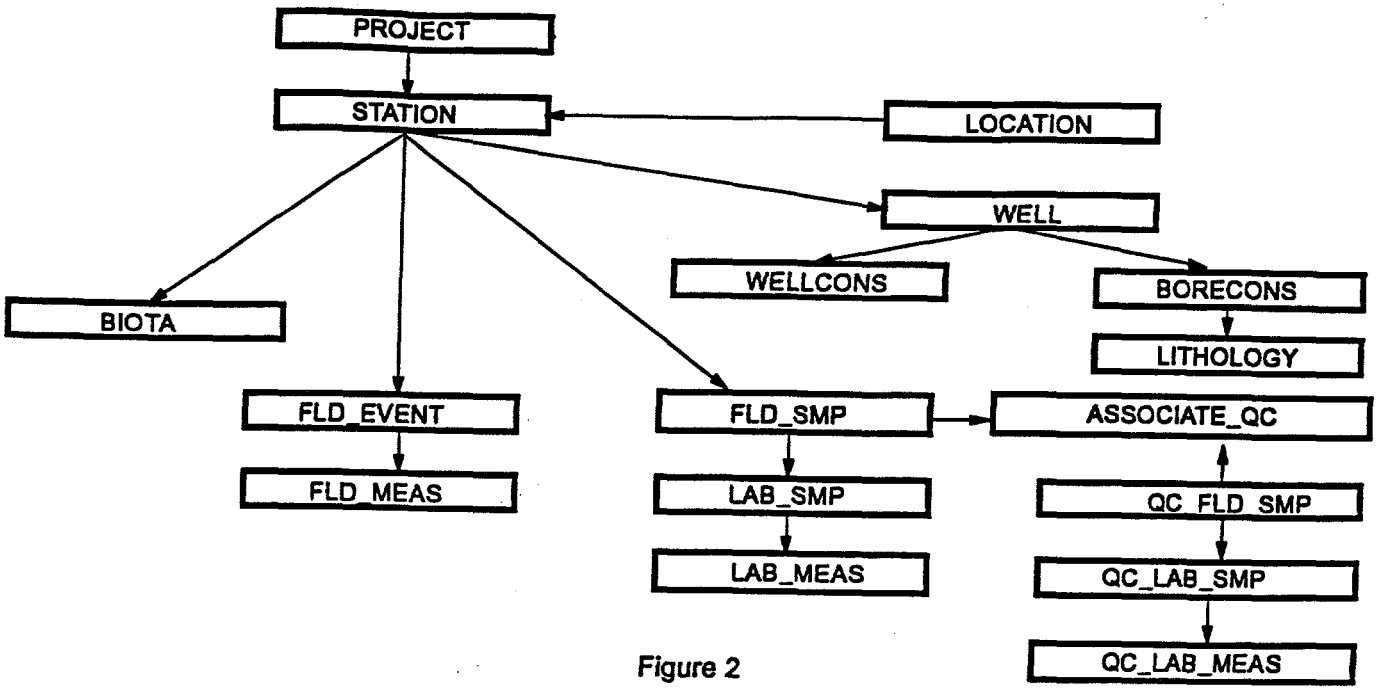

Figure 2

PRIMARY AND FOREIGN KEYS

Primary keys in all the tables except the reference tables and ALIAS_SOURCE are system-assigned sequence numbers. These keys are all in the format *table_name*_ID. For example, the primary key of PROJECT is PROJECT_ID. (The exception to this rule is TRANSMITTAL, whose primary key is TRANS_ID.) These keys are used to tie the tables together, because OREIS cannot count on getting unique identifiers from the projects. All primary keys are automatically indexed by ORACLE. OREIS defines additional indexes for the fields on which queries to the database probably would be based.

A number of foreign key integrity constraints have been established among the OREIS tables. These constraints are used to enforce parent/child relationships. (See Figure 3.)

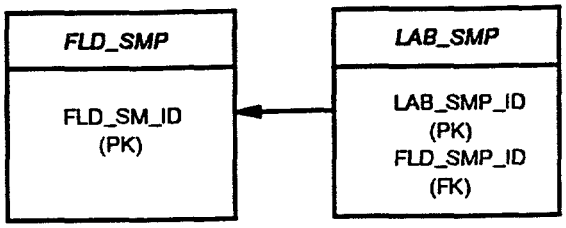

Figure 3

PROCESSING THE DATA

The OREIS database manager receives a transmittal package from the project. This package includes an OREIS Transmittal Form, any relevant documentation, the data in an electronic format, and frequencies and contents listings of the data. The database manager gives the transmittal package to the configuration control specialist to initiate the processing. The specialist reviews and logs in the package. In addition, the specialist ensures that the appropriate records are entered in the PROJECT, DATASET, and TRANSMITTAL tables. When the configuration control specialist has finished these tasks, the database manager assigns the data package to a staff member for processing.

Most of the data arrive as SAS data sets. If they do not, they are converted into SAS format. (For example, data in Xbase format can be converted to SAS data sets with PROC DBF, available with the SAS System for PCs). All of the initial data processing is performed in SAS.

The first part of the data processing is a manual mapping of the OREIS fields with the SAS variables. The OREIS ALIAS_SOURCE table is used to capture this mapping. It contains the table

and field names of the OREIS tables and the associated data set and variable names from the source data. The OREIS data processors are in frequent contact with the project data generators during this phase to ensure that the OREIS interpretation of the project's data is accurate. In addition, the data processor checks for possible data corruption by checking record counts and summary statistics against values provided in the data transmittal package.

Because of its relationship to most of the measurement tables, STATION (and, therefore, its parent LOCATION) must be processed and checked before any other data processing of the transmittal can occur. The SAS processing program that populates these two tables must check that coordinate information is entered for each record, either northings and eastings or latitudes and longitudes. Only one record per unique set of coordinates is permitted in the OREIS LOCATION table; however there may be many STATION records for each LOCATION record. Checks of this integrity constraint are carried out separately in a GIS system.

A unique station name must be included in the data, as well as the type of monitoring activity that occurred at the station. If more than one type of activity took place at a given station, multiple records will be created for the OREIS table.

Unit and code conversions are a part of the SAS processing. Code verification is accomplished by using SAS/ACCESS and PROC SQL to (1) bring valid OREIS codes into SAS data sets and (2) find any values that do not match OREIS codes for the given variable.

The following is a portion of a SAS program used to compare the parameter values in the OREIS PARAMETER table with the parameter values in the dataset being processed, SURFWAT. The SURFWAT dataset contains the columns PARAMTR, which is null for each record, and CAS_NO, which has a value for each record. The ORACLE table OREIS.PARAMETER contains the columns PARAMTR and CAS_NUM, which have no null values.

The first step is to create a SAS dataset from the ORACLE table using SAS/ACCESS. This step utilizes the user name and password from the "OREIS Operating Environment" section of this paper. Step 2 is a simple SQL update statement that uses the dataset created in step 1 to update the SURFWAT dataset. Step 3 creates a SAS dataset that contains all the records from SURFWAT that do not match the records in the ORACLE PARAMETER table. These records will be assigned pseudo-parameter values later in the program.

```
Step 1
proc sql;
connect to oracle(user = &user
   orapw = &password
   path = &twotask);
create table paramter as select * from
   connection to oracle
   (select cas_num, paramtr from
    oreis.parameter);
quit;

Step2
proc sql;
update surfwat
   set paramtr = (select paramtr from
   paramter where cas_num = cas_no);
quit;

Step3
proc sql;
   create table noparm as
   (select cas_no from surfwat where
   paramtr is null);
quit;
```

Throughout the entire data processing procedure, the data undergo review for consistency and completeness. This review includes (1) scrutinizing frequency tabulations and sorted lists of units, qualifiers, codes, and other selected character fields to check for missing values, miscoded data, and inconsistencies; (2) comparing new and existing data to check for errors and inappropriate units of measure; (3) examining range checks, elementary statistics, or scatter plots of numeric and date fields to check for missing data, unreasonable values, and outliers; (4) exploring maps of station locations to verify coordinates; and (5) evaluating other checks based on the reasonableness of the data (e.g., checks for samples from dry wells).

Data associated with each OREIS table are checked using specific instructions that have been written for this purpose. One example is a check to

see that the date a project is completed is after the date the project was initiated. Other checks include frequencies of character variables (such as sample methods, sample types, and chemical names) and statistics of numeric variables (such as minimum, maximum, mean, and standard deviation). During the data processing problems may be encountered, such as missing values in required fields, project codes that do not easily convert to OREIS codes, and missing station coordinates. Problems are resolved with the assistance of the OREIS data coordinator and, if needed, the project data custodian or project data generator. The data conversions and problem resolutions are documented in the OREIS data processing report.

As one of the final pieces of the SAS processing, values are assigned to the primary key variables which contain system-assigned sequence numbers. These numbers, which are generated by a separate program that accesses ORACLE sequences, are then coded into the SAS program. Even though this is not the most effective way of doing this assignment, it was chosen to ensure an unbroken series of unique numbers for each set of data.

The last step in the SAS processing is to create SQL*Loader files which are used to move the records into the TRANS tables, where they will remain during the entire QA and data package approval process. (SQL*Loader is an ORACLE utility.) As mentioned earlier, there are no foreign key integrity constraints on the TRANS tables. The primary key constraints ensure that no records with duplicate keys can be loaded.

The last step in the entire data processing procedure is the movement of the approved data from TRANS to the OREIS tables, where they become accessible to users. The program used to move the data was written using Pro*C, an ORACLE precompiler that allows SQL code to be embedded within a C program.

SAS/ACCESS and PROC SQL

SAS/ACCESS and PROC SQL are used extensively in the processing of the data and in the QA of the records once they are put into the TRANS tables. In the processing phase, parent tables are brought into SAS and used to update foreign keys. In the QA programs, data sets are created from the TRANS tables and then checked for valid codes, date consistency, and outliers. Summary statistics are generated from these data sets and compared with statistics delivered as part of the data transmittal to validate the data processing.

CONCLUSION

SAS/ACCESS and PROC SQL are useful in large database applications. They provide the benefits of relational database structure and SQL syntax, while retaining the statistical and analytical capabilities of SAS software.