CONF-950868--34

Control #1206
Beverly Cather Zygmunt
Oak Ridge National Laboratory[1]
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

A Consolidated and Standardized Relational Database for ER Data

The three U.S. Department of Energy (DOE) installations on the Oak Ridge Reservation (ORR) (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. Lockheed Martin Energy Systems manages the three Oak Ridge installations (as well as the Environmental Restoration (ER) programs at the DOE plants in Portsmouth, Ohio, and Paducah, Kentucky). DOE Oak Ridge Operations has been mandated by federal and state agreements to provide a consolidated repository of environmental data and is tasked to support environmental data management activities at all five installations. The Oak Ridge Environmental Information System (OREIS) was initiated to fulfill these requirements. 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.

This paper discusses the benefits of a consolidated and standardized database; reasons for resistance to the consolidation of data; implementing a consolidated database, including attempts at standardization, deciding what to include in the consolidated database, establishing lists of valid values, and addressing quality control (QC) issues; and the evolution of a consolidated database, which includes developing and training a user community, dealing with configuration control issues, and incorporating historical data. OREIS is used to illustrate these topics.

What is a Consolidated Database?

A consolidated database brings together and unifies dissimilar data. According to the dictionary, consolidation is "to combine into one, to make strong and stable." This combination may be the result of information stored in different formats and media, data that were collected for different purposes, or data maintained at different sites. A consolidated database can be a database of environmental measurements located on separate but networked machines; a collection of sampling information files, electronic or otherwise (e.g., log books), or pointers to files; tables

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

Control #1206
Beverly Cather Zygmunt
Oak Ridge National Laboratory
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

with disparate or similar structures in project-level database management systems; or raw measurement data.

The definition of consolidation includes becoming "strong and stable." A consolidated database gains strength and stablity as it encourages the use of a single set of tables, files, or pointers rather than permitting mutiple copies of data, which are difficult to maintain. But the true strength of most consolidated databases comes from the standardization of the data. In an environmental measurements database this standardization may include coded chemical names, standard date formats, and specific required fields.

In OREIS, this consolidation involves the joining of tabular data from sources such as environmental restoration projects and compliance programs. These data traditionally have been stored in a variety of electronic formats or as log-books and other "hard copy" materials. In both cases, but particularly the latter, the data could be viewed only by a very limited number of persons. OREIS makes these data available, in a standard format, to a much larger user community.

Benefits of Consolidation

On the ORR, the consolidation of the data is especially important because of a move towards thinking of the three separate plants as a single unit. Having the data on-line in a single relational database management system supports this new initiative. The various projects and historical data sources now are required to standardize their data to fit into OREIS, and this standardization gives data for the entire reservation more of the same "look and feel" to the users of the data.

Another reason to implement a consolidated database, especially in the current trend towards downsizing, is economics. It is more expensive to maintain duplicate sets of data than a single, unified data repository. Additional resources must be committed to keep the data consistent when they are stored in separate systems. Costs associated with data storage include hardware and software purchases and maintenance, but the most significant expenditures are for the personnel needed to create, update, and maintain these redundant systems.

A benefit of consolidation is an improvement in the quality of the data. OREIS checks for anomalies that may be found when data from one project are compared with data from another.

It is easier to use the data if they are in one place and of known quality. Even within projects, there was often a lack of standardization and consolidation. This absence of unifying

Beverly Cather Zygmunt
Oak Ridge National Laboratory
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

characteristics made it harder for projects to share data, required reports to be rewritten for each project, and increased the cost of subcontracting the work. When a user needs to access from more than one project, it is less costly when the data are in a single repository.

Resistance to Consolidation

Despite the benefits, there is often resistance to the consolidation of data. Most people are initially opposed to changes in the way they do business, especially if they feel their input has not been given enough consideration. In OREIS, the data custodians were unable to have their opinions about the need for a consolidated database considered, because that decision had been mandated. This circumstance created initial resistance to the consolidation effort.

In addition, custodians often feel ownership of their data. This ownership is expressed by a desire to maintain control of and access to the data. Historically, when projects were smaller and their associated data did not need to be shared with other projects, this was the way business was done. That viewpoint, however, is not consistent with the consolidated approach.

Another reason for resistance is the fear that consolidation will lead to downsizing/rightsizing, with fewer people needed to work with the data. Some of this work, which includes updating and maintaining the data and answering user requests, will be taken over by the consolidated database staff and will mean fewer jobs at the project level. The data custodians will be responsible for gathering and manipulating the data to send to OREIS, but some project personnel may be reassigned to the central staff.

Although saving money is a benefit of consolidation, the costs to the projects of "retrofitting" the data to match the consolidated database structure can be substantial and a reason to resist. This retrofitting includes format changes, code revisions, keeping data that were never kept before, and putting data into an electronic format. Retrofitting takes time and, therefore, costs money. At OREIS, because historical data are more expensive to retrofit, the data are subjected to less stringent requirements than the data from current projects.

Once the decision has been made to implement a consolidated database, standards must be created, viewpoints must be considered, and there must be a "buy-in" from the people who will be sending the data. OREIS has managed to overcome many of the objections by showing data custodians and project staff how the implementation of a consolidated database can be used to benefit them by reducing data management tasks, including responding to user requests; defining standard codes and formats; and providing a larger context in which consistency checks of the

Beverly Cather Zygmunt
Oak Ridge National Laboratory
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

data can be performed. An unexpected result of OREIS's work is that ownership of the data has been redefined. Although initially there was resistance to losing control of their data, now data custodians consider themselves and their data part of the OREIS system.

Standards

A variety of standards must be established before a consolidated database can be implemented. A set of valid values, or codes, is used to standardize the data. OREIS reviewed codes used by existing environmental programs and databases. Committees on the ORR that were already developing common practices were also consulted for their rules and lists of values. Based on this information, OREIS established its own set of codes. These codes were reviewed internally to ensure that they followed standard naming conventions. OREIS continues to work with data users and experts in various ER-related fields to develop new codes, as needed, and refine existing codes.

Another standard that must be put in place is the definition of what goes into the consolidated database. A decision was made that OREIS would include the following: only data that have been cleared for public release (i.e., no classified data); summarized data, when appropriate (e.g., flow readings); and data that are required by the Environmental Protection Agency (EPA) and DOE. Because some transformation of the data may occur in the processing to comply with OREIS standards, the processing must be approved by data generators/custodians before the data are loaded into OREIS.

A standard means of defining locations was required because there are five different northing/easting coordinate systems (K-25, Y-12, ORNL, Administrative Grid, and State Plane) on the ORR used to locate sampling stations. In an attempt to standardize, it was decided to make Administrative Grid the official OREIS coordinate system. All coordinates are translated to this system. Because delivery to the EPA must be in latitude/longitude coordinates, OREIS also generates and stores these.

Standardizing units is a goal towards which OREIS is working. OREIS had to revise its initial list of valid units for new data sources. OREIS is leading the effort to define a standard unit for each unique combination of analysis type, analysis method, and analyte.

The original list of valid analytes was obtained from an internal standards committee and was based on the Chemical Abstract Services Registry Numbers (CASRNs). Many analytes of interest to OREIS were outside the scope of this committee and did not appear in the original list. OREIS

Beverly Cather Zygmunt
Oak Ridge National Laboratory
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

followed the format of the committee's list when adding analytes and used CASRNs when available, but in some cases had to generate nonstandard identifiers. Periodically, the OREIS analyte table is reviewed to determine whether multiple listings of an analyte, due to chemical synonyms, have occurred.

The OREIS methods have a similar history. They began as a list obtained from a standards committee. When required by new projects, OREIS added methods, following the original format. OREIS currently is working to standardize these methods.

QC Issues

In a consolidated database of ER data, a decision must be made as to which, if any, QC results are stored. Currently, OREIS maintains only the field QC samples, such as rinsates, field blanks, and trip blanks, that are actually sent to a lab and are associated with one or more measurement samples. An ongoing question is whether OREIS needs to keep any lab QC data, such as lab blanks and matrix spikes, which are used to calibrate the lab equipment. Potential users and risk assessors may need these data in order to determine usability. Although some projects consider field duplicates as QC samples, OREIS considers a field duplicate as a regular sample and treats it just like any other sample.

One of the primary mandates of OREIS is to store data of "known" quality and make those data readily available to ER Program professionals, regulators, and the public. Originally, the OREIS policy was not to store data that had been qualified as "rejected" during the validation process. These data were assumed not to have been used in making decisions regarding a given site and there was the potential for using a "bad" data result out of context and drawing an erroneous conclusion. However, several issues have arisen over time that resulted in OREIS storing some "rejected" data. These issues are as follows.

- OREIS is mandated to keep data that are used in a Federal Facilities Agreement deliverable or are used to support the conclusions contained in the report; the "rejected" data may have been used to support some portion of the report.

- It is possible for data to be rejected by a project yet still be useful to secondary users. Because of this, the ER community has moved away from strict validation based on rigid guidelines and has adopted the Data Quality Objectives (DQO)-based approach for determining data usability. The Data Quality Program at ORR will accelerate work to

Control #1206
Beverly Cather Zygmunt
Oak Ridge National Laboratory
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

define and implement a standard set of usability qualifiers that will replace the currently accepted validation qualifiers.

Standard Formats

Initially, OREIS would accept data in any electronic format. This decision was made to encourage people to send data to OREIS. Even though this approach made sending data easier for the data generators, it made OREIS processing very difficult. Now, OREIS only accepts data in .DBF, SAS® export, ORACLE® export, and ASCII.

OREIS Ready-to-Load Format

The first data packages received by OREIS took months to process. Many problems were encountered in trying to map the data into the OREIS structure: fields that had the same name but meant something different; fields with dissimilar names that meant the same thing; analytes and analysis methods that did not match the OREIS list of valid values for a variety of reasons, including spelling, capitalization , synonyms, and revision numbers; missing values for required fields; QC samples that were difficult to match to regular samples; and inconsistencies within the dataset, such as outliers, unit conversion errors, duplicate records, and missing values.

A recent ER management decision was made that the projects should bear more of the cost of getting data into OREIS. Most of the responsibility of preparing the data to OREIS specifications has now been passed to the projects. An OREIS ready-to-load (RTL) specification gives projects guidance on how to do this. The RTL format includes lists of codes, descriptions of the OREIS fields, file structures, and a model for linking QC and regular samples. Data provided in RTL format can be processed in weeks, rather than months, because there are fewer issues to resolve: the project has already made the decision about how to map fields, code conversions have been made to comply with OREIS standards, values have been assigned for all required fields, and the link between QC and regular samples has been defined. Problems may still exist with (1) the data, such as outliers and errors in units conversions; and/or (2) the need to add additional codes.

Evolution of a Consolidated Database

As a consolidated database evolves, a number of issues arise. When new data are submitted, changes may be required: changes to the code tables, the model, and documentation. Any of these changes may require a retrofit. These changes must be communicated to the data generators and contractors, and may affect the way in which data generators do business.

Control #1206
Beverly Cather Zygmunt
Oak Ridge National Laboratory
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

As the database and number of users grow, performance tuning becomes crucial. Hardware and software upgrades may be required to improve database response and to provide for more users of the system.

As more data become part of the consolidated system, there is even less reason to keep redundant copies of the data at the site or project level. It also becomes important to provide a usability flag that defines usability more broadly than the project data qualifiers.

The users' experience and expectations grow as new data become available to them and they gain familiarity with the system. These expectations help provide direction for growth of the database and new functionality. Data generators are identifying new sources of data for OREIS, including more historical data than they initially planned to send.

On the ORR, a trend towards becoming "leaner and meaner" has emerged at both the projects and OREIS. A pilot project geared towards rapid sample turnaround is under way. OREIS data processing staff are periodically updating procedures to provide a more timely processing of data.

Developing a User Community

Initially, the OREIS user community was defined by the federal and state mandates to create a consolidated database. This community included regulators at both the state and federal level and DOE. The user community has grown. Recently, OREIS provided data to two out-of-state groups doing dose-reconstruction projects to re-evaluate the contaminant exposure to people who worked on or near the ORR twenty years ago. These groups came to OREIS because OREIS could provide a variety of data in electronic format.

OREIS is continuing to expand its user community as more and new types of data are loaded. New types of data that will be coming to OREIS include risk models, additional types of compliance data, decommissioning and decontamination (D&D) data, additional historical data, and United States Geological Survey (USGS) and National Oceanographic and Atmospheric Administration (NOAA) data. Users can access data from a variety of projects and, because of the standardization of the data which has taken place, the same variables, codes, and definitions are found in each. The National Pollutant Discharge Elimination System (NPDES) and ambient air compliance data are being supplied regularly to OREIS. As a result, these groups will generate their parts of the next ORR annual report from data within OREIS. In addition, current projects are generating reports from the data in OREIS, using both project-generated and historical records.

Control #1206
Beverly Cather Zygmunt
Oak Ridge National Laboratory
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

An educated user community is a requirement for a successful consolidated database. One of the things a user must know is the relationship among the data fields. For example, in OREIS the user must examine both the result and result qualifier fields when evaluating results because the qualifier can indicate that a given result was rejected. Similarly, in the case of NPDES data, if the significant digit value is not applied to the result, it can appear that a result is out of compliance, when it is not.

In addition, users need a basic understanding of the data model and, since many of the fields are coded, a general understanding of the types of fields that are coded and the code values. Educated users are an asset because they are are able to envision uses beyond the system's current capabilities and contribute to the growth and direction of the consolidated database.

Configuration Control

A consolidated database, as it adds data sources, is by its very nature a moving target. As new types of data are brought into the database modeling enhancements may need to be made. Code lists may need to be expanded. The opportunities for developing new applications increase, and current applications may be need to be modified. Configuration control, performance tuning, and backups, which are always an issue, become even more critical. The documentation must be kept current. There may also be a need for additional procedures and instructions as new data sources are added. In a rapidly evolving system, it becomes more of a challenge to keep the users updated on all of the changes. OREIS is moving towards making its documentation available on the World Wide Web to meet this challenge.

A part of configuration control is the need to retrofit data to reflect changes to the model. All retrofits involve basically the same steps:

- Summarize what is currently in the database. The statistics generated in this step will be used to verify the transformation of the data.
- Determine how to convert data from the old structure to the new.
- Write programs to move data from the old structure to the new.
- Move a test set of data into the new model.
- Verify that the transformation of the test set is correct.
- Move all of the data into the new structure.
- Verify that the transformation is correct by using the statistics from the first step.

Control #1206
Beverly Cather Zygmunt
Oak Ridge National Laboratory
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

As the database becomes larger, this process becomes more time-consuming, even for simple changes to the structure.

Incorporating Historical Data

Historical data often do not meet current database specifications. Required fields may not be available. Code definitions may be missing. Even data that were collected six months ago may not meet today's standards. But there is a wealth of information in historical data, and it is worth the effort to accommodate these data.

Another issue that arises with historical data is how to deal with duplicate records. If two projects pulled information from the same source and both projects used these data in a report, both projects are required to send these data to OREIS. Because OREIS does not want to store duplicate measurement records, changes to the model have been made so that a single result can be associated with multiple projects. In this type of structure, in which a record can be "owned" by more than one project, configuration control issues become complex.

Conclusion

Considerable effort is required to implement and to maintain a consolidated database, but the benefits outweigh the costs. Resistance to consolidation must be overcome. Standards must be established and updated to accommodate new data types. In a consolidated database secondary users must be able to determine the quality of the data. Acceptable formats must be specified, such as date and time conventions and the file type used to transmit the data (e.g., SAS® export). In order to facilitate data transfer and data processing, it is beneficial to define a structure for data transfer, including field names, valid codes, field types (e.g., character), maximum field lengths, and required fields.

As the consolidated database evolves, a number of issues arise, including changes to the code tables, model, and documentation, which makes configuration contol a primary concern. Performance tuning becomes crucial as the number of records and users of the database grows. Evolution may be directed by users as they become more familiar with the system functionality and the contents of the database. Increased functionality and content lead to even more users.

Benefits of a consoliated database include reduced costs, because it is no longer necessary to maintain redundant copies of the data. Standardization of the data improves usability, and a check for anomolies improves their quality. Consolidation allows users to access data from a variety

Control #1206
Beverly Cather Zygmunt
Oak Ridge National Laboratory
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-574-1007 FAX: 615-241-5324

of projects and, because of the standardization of the data which has taken place, the same variables, codes, and definitions are found in each.

Co-Author: Teresa James
University of Tennessee, Knoxville, TN
P.O. Box 2003, MS 7056
Oak Ridge, TN 37831-7056
Phone: 615-241-4881 FAX: 615-241-5324