XIX International Conference on Water Resources
CMWR 2012
University of Illinois at Urbana-Champaign
June 17-22, 2012

A METHODOLOGY FOR MANAGEMENT OF HETEROGENEOUS SITE CHARACTERIZATON AND MODELING DATA

Deb Agarwal[1], Arthur Wiedmer[1,2], Boris Faybishenko[1], Tad Whiteside[3], James Hunt[2], Gary Kushner[1], Alex Romosan[1], and Arie Shoshani[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720
Computational Research Division and Earth Science Divisions
1 Cyclotron Rd, Berkeley, CA 94720
e-mail: DAAgarwal@lbl.gov, BAFaybishenko@lbl.gov, AShoshani@lbl.gov, ARomosan@lbl.gov, and GEKushner@lbl.gov

[2]University of California, Berkeley, CA 94720
Department of Civil and Environmental Engineering
e-mail: arthur.wiedmer@berkeley.edu, hunt@ce.berkeley.edu

[3] Savannah River National Laboratory, Aiken, SC 29808
e-mail: Tad.Whiteside@srnl.doe.gov

Key words: Data management, ASCEM, Savannah River, Google maps, data intensive

Summary

Scientists, site managers and regulators participating in environmental remediation activities usually need to access, create, and manage large amounts of heterogeneous data, including disparate site characterization data, and models that range from conceptual to numerical. The data are necessary to enable the development of conceptual and numerical models, parameter estimation, and multiple simulations. We designed a methodology based on DOE Savannah River Site F-Area and Hanford databases that included well layout, concentration, groundwater level, lithology, meteorological records, contaminant levels, etc. The approach was necessary to effectively manage the staging and views of environmental data to ensure that site characterization and monitoring data are readily available as inputs for parameter estimation, numerical predictions, uncertainty quantification, and risk analysis. This paper describes the methodology and the application of the developed data management system to the Savannah River F-Area site observations. In particular, we present methods for organizing different databases and data types into a common framework that allows the user to browse the data as a single coherent dataset. Analyses of the Savannah River F-area using the data management system illustrate the utility of the system. The paper also demonstrates the utility of interfaces designed to leverage and integrate existing large-scale data management, provenance and indexing technologies from the scientific community and discuss the next steps.

1 INTRODUCTION

The last 3-4 decades of extensive subsurface field monitoring, site characterization, and numerical modeling at the U.S. Department of Energy (DOE) Environmental Management (EM)

cleanup sites have generated tremendous volumes of data. A new toolset to integrate disparate databases, transparent to the users is needed to enable broad usage of these data. Moreover, the user may need to collect data from different databases (e.g., relational, hierarchical, spreadsheets) and sources, each having their own infrastructure. The dispersed and heterogeneous nature of the datasets, which is common to many other contaminated DOE sites, hinders effective use of the data for advancing the DOE EM cleanup efforts.

Advanced Simulation Capability for Environmental Management (ASCEM) is a DOE EM project building the integrated set of software components and models needed to support a scientific approach to understanding and predicting contaminant fate and transport in natural and engineered systems. The resulting modular and open-source high-performance computing tool will facilitate integrated approaches to modeling and site characterization that enable robust and standardized assessments of performance and risk for EM cleanup and closure decisions (http://ascemdoe.org/).

As part of the scope of the ASCEM program, the Computer Research Division of LBNL (jointly with Earth Sciences Division of LBNL, University of California, Berkeley, and SRNL) has been involved in the development of a new state-of–the-art database and data management system, using as a case study the Savannah River Site (SRS) F-Area having site characterization and long-term monitoring data. The F-Area datasets are distributed among multiple spreadsheets, scientific reports and publications, databases, and data files, all stored in different formats and on different computers. The ultimate goal of the ASCEM Data Management component is the development of an integrated knowledge management environment that enables users to easily find, access, and contribute then utilize the combined knowledge and data stored in the system for synthesis.

This paper provides some background about the types of data available from the DOE cleanup sites, presents the methodology of the data management system, and describes a use case to demonstrate how the data can been used to understand the inventory of subsurface contaminants.

2 DOE ASCEM CLEANUP SITES AND DATA

The Hanford[1] and Savannah River[2] Sites were the U.S.'s nuclear production facilities. The Hanford Site was established in 1943 for the production of plutonium for the Manhattan Project and was expanded during the Cold War to continue plutonium production and processing. Over 900,000 m^3 of liquid and solid radioactive waste stored at the Hanford Site requires disposal. Construction at the Savannah River Site began in 1950 for the production of nuclear materials used in nuclear weapons. Some of the low-level radioactive wastewater from the separation facilities was disposed of in the F-Area seepage basins, assuming that the underlying sediments would to contain the waste and delay its seepage to surface waters. These basins and the surrounding area are under study for the ASCEM demonstration project.

The ASCEM models require a variety of data inputs. The sources of this data are located in multiple formats and are "owned" by multiple organizations. The formats include paper-reports, spreadsheets created by project-leads, individual databases (MS Access), and enterprise class databases (Oracle). The lack of consistent structure is primarily due to two factors: age and

diversity of sources. The first environmental and geologic data was acquired in the 1950's; these data have undergone system changes, multiple owners, multiple users, additions, and deletions.

The data used in development of the data management system includes site-wide information, such as meteorological and geological conditions, operational data such as well pumping schedules, and remediation action data such as base injection operating parameters. While the site-wide data are comprehensive, they are owned by different departments and kept in formats relevant to those organizations. So, the site managers and scientists needed to apply significant efforts to find and collect data from multiple sources into a single unified place, format, and units. Before the data management system, users who wanted to synthesize information from multiple sources of information, spent extensive effort to find and collect data into a single unified place, format, and units. The system and data used as described below.

3 ASCEM DATA MANAGEMENT SYSTEM

The overall objectives of the development of the methodology for the ASCEM data management system are: 1) to provide a data and information infrastructure that is accessible to all the components of ASCEM, including parameter estimation, numerical modeling, conceptual model development, etc, and 2) to enable users to easily access, browse, and download the data. To ensure consistency, reproducibility, and traceability of ASCEM analyses, it is critically important to effectively manage the staging and views of data, and to ensure that input data are ready and available to a computation along with the corresponding metadata, version, and provenance information.

3.1 System Design

The Observational Data Management System (ODMS) was designed with two goals: (1) provide a way to ingest data from heterogeneous data sources into a single database schema, and (2) generate common data views for driving a generic interface for data browsing, searching, display, and output. These goals avoid the need for special purpose database schemas and user interfaces for each new database, and ensure a unified management and view of heterogeneous datasets.

The data management modular architecture components are shown in Figure 1. The system design begins with the development of the *staging system*. Staging includes ingesting the data into the database, cleaning the data, standardization of variable names, and any unit conversions required. Once the data have been staged, they are moved to the *archive database*. The archive database stores the data in a uniform format, allowing new data types to be incrementally added

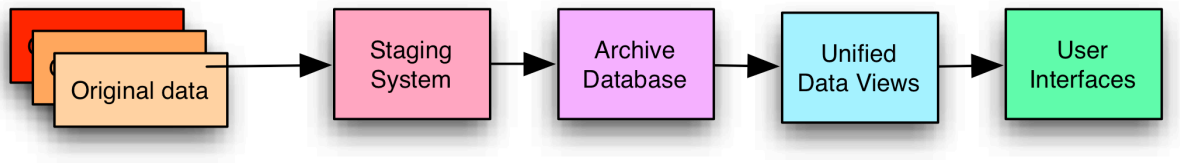

Figure 1: ASCEM data management system architecture view.

to the database without requiring major redesign of the database schema. We use a canonical schema, where common concepts such as station/well and types of data (e.g. analyte concentration and lithology data) are unified into separate tables from the data table. The *unified data views* are instances of the data organized in the ways that the user would like to access the data. For example, a data view may reconstruct the time history of contaminants in the wells. The *user interfaces* utilize the unified data views to present data processing, access, and browsing interfaces.

3.2 System Implementation

A schematic of the system implementation is shown in Figure 2. In the staging process, the original data are ingested, cleaned, and converted into standardized "cleaned templates." This process is now automated for data formats we have received to date from the SRS. The current cleaning process is fairly crude, and continues to be refined over time as new data issues are identified. A more extensive cleaning process is planned as a future task. Once the cleaned templates are generated, they are annotated and ingested into a SQL Server database system. The archive database step is an automatic "shred" of the data into the canonical schema. From that schema, several views are generated. These views are designed for automated display of the data through user interfaces, as shown at the bottom of Figure 2. Thus, the entire process of populating the database and preparing the data for user interface display is automated.

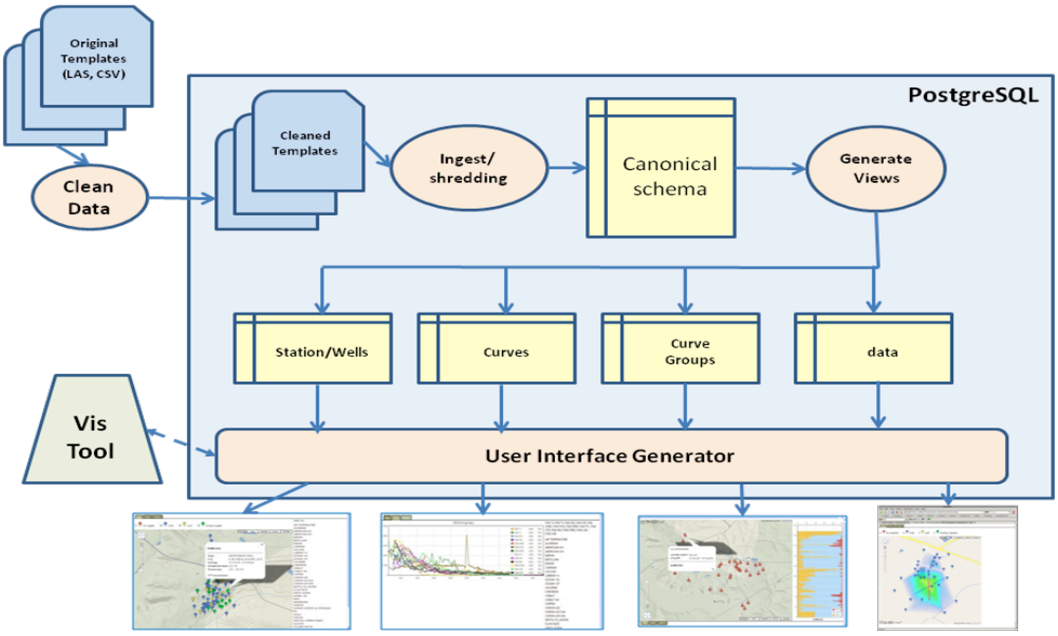

Figure 2: Schematic of the data management system implementation

Data Staging System

The majority of the SRS data came from two databases. Below is a description of these sources.

- Landmark depositional database[4], contains information about lithology, well coordinates, well depths, well-screened zones, depths of stratigraphic units, particle size distribution, and depositional environments, as well as cone penetrometer (CPT) data. These data were received as Log ASCII Standard (LAS) files.
- BEIDMS (Bechtel Environmental Integrated Data Management System) concentration database, including measurements of concentration of 44 analytes and other wellbore parameters collected over 1990–2011 from 145 monitoring wells. These data were received in the form of an Access database.
- Hydrostratigraphic database, provides coordinates of the base of the hydrostratigraphic units determined from wellbore data[4]. These data were received as an Excel spreadsheet.
- GIS data, which include digital elevation models (DEM) from topographic maps.
- Hydrologic data, including hydraulic properties of saturated sediments, extracted from a large-scale flow model[4,5], and unsaturated soil properties (water retention and relative permeability curves)[3]. These data were received as Excel spreadsheets and reports.

The information stored in the developed data management system include: 9 abandoned extraction wells (71,061 data points), 11 abandoned injection wells (130,846 data points), 21 abandoned monitoring wells (105,113 monitoring wells), 3 characterization boreholes (101,427 data points), 3 composite wells (58,903 data points), 121 CPT-U (cone penetration tests tip resistance, sleeve friction, and pore water pressure measurements-588,894 data points), 35 monitoring wells (330,688 data points) and 18 PCPT (piezocone penetration test) wells (26,055 data points). The developed data management system also includes meteorological data, contaminant source release data to the F-area basins, and a series of relevant environmental and other reports.

The datasets stored in the ASCEM Data Management system are categorized into two groups: (a) measured or simulated data, referred to as “transparent data”, and (b) documents, graphs, pictures, and similar data objects, referred to as “opaque data.” Transparent data can be searched and extracted while opaque data can only be accessed as a whole, although text documents can be indexed with keywords. This paper describes the transparent data system, also referred to as the Observational Data Management System.

The data staging process for each of these data groups above starts with the conversion of the received data into a format the staging system can ingest (as described in Section 3.1). The data are then cleaned to handle data that are not measured values (e.g. containing ~, <, >) and data types that are not standard (naming different from the standard variables or a data type not previously encountered).

Archive Database

The archive database relies on a schema that has been vetted and is in use on a wide variety of projects, including the California watershed server, FLUXNET, and the National Soil Carbon Network databases[6,7,8]. This star schema utilizes a separate table for data type, site/layer, and time. Each of these tables has a primary key. The data values are all stored in a single data table with one data value per row indexed by keys from the other tables. We refer to this as a shredded schema, because it breaks down all the original big wide tables provided from the various

sources.. Additional columns in the data table allow one to track provenance information, data version information, and any additional modifiers related to the data. This schema has proven robust to changes in type and quantity of data. For more information about the schema, see the references provided earlier.

Data View Tables

The archive database provides a long-term storage format that is easy to maintain. However, its organization is inefficient for answering the types of questions and providing the types of data views that a user typically would like to see. The data view tables are automatically generated in the database and provide the reconstructed views of the data, which support the ways users would like to operate the data. The advantage of storing the data in the shredded database and then creating the data views is that these user data views can change very easily, since they are created on the fly. Requests to change the way that the data are viewed are easily accommodated without changing the data storage structure.

User Interface

Two principles guided the development of the user interface: (1) driving the interface from canonically generated views, using the same views for all data; and (2) the use of open source software. We selected PostgreSQL as the database system for the ingested data and for generating the views. The requirements for the interface included the display of wells/stations on a geographic map. For this purpose, we used "Google maps" and adapted the display/legend to show wells/stations by type, using the icon shape and color. When clicking on a well of interest, we used the display capability of "Google maps" to show summary information, such as the well coordinates, types, etc. At the same time, we used Javascript to retrieve from the database (automatically generating SQL queries), and to display a list of analytes (that have time series measurements) or depth elements (such as moisture content). Clicking on an analyte name, brings a temporal display, using the open source plot tool--called FLOT, which is also used for the depth elements. Finally, a "data" tab shows the displayed data in a tabular form, and a "save" button saves the result to a file to be used by other scientists involved in the ASCEM project, such as inputs for simulation runs. Some examples of the user-interface displays are shown at the bottom of Figure 2. The current ASCEM data user interface is available at http://babe.lbl.gov/ascem/maps/SRDataBrowser.php. A tutorial describing the operation of the interface is available under the help button.

4 EXAMPLE USE CASE

The current version of the ASCEM data management system is already in use to support the development of input parameters for modeling contaminant transport at the SRS F-Area. The work to estimate the groundwater tritium inventory at the Savannah River F-area used historical groundwater and surface water monitoring data. Tritium was selected as the most frequently measured radioisotope and it serves as a chemically inert tracer. The goal was to establish the tritium inventory in the subsurface over time by means of an overall mass balance using the reported annual load to the basins and estimated releases to Four Mile Branch, a creek located

down gradient from the basins. The historical groundwater monitoring record provides an alternative means to check the amount of tritium in the subsurface over time.

The ASCEM web interface was used to determine wells needed to delimit the plume. We used On-Line Analytical Processing (OLAP) capabilities to pre-compute yearly averages of tritium activities at every well. We also used water levels and aquitard (aquifer boundary) elevation data. The data were imported into MATLAB through the datacube OLAP interface. Using MATLAB the tritium activity was interpolated in the uppermost aquifer on a regular grid. Total tritium was then computed through a discrete integration.

Results in Figure 3 demonstrate an agreement within a factor of two between these separate and independent means of calculating tritium subsurface inventory, with red lines derived from an overall tritium balance and the violet curves representing an inventory determined from monitoring well data interpolation. The decline in tritium inventory fits an exponential decay model quite well with the decay constant being the sum of the flushing time for the aquifer and the tritium radioactive decay constant. An application of the exponential decay model for a monitoring well FSB 78, located next to Basin 3 can predict when drinking water standards are expected (Figure 4).

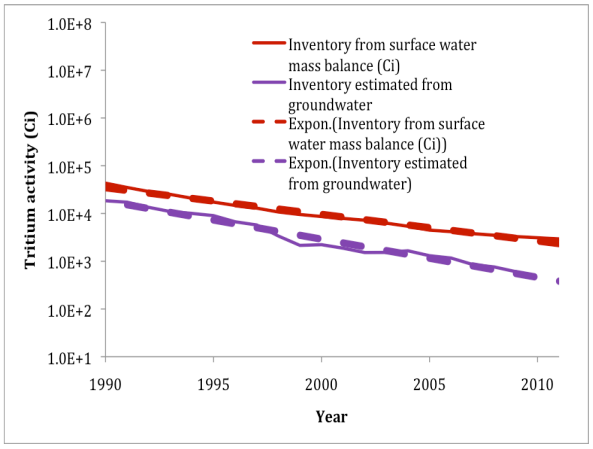

Figure 3: Tritium inventories computed using source term and surface water data (red), and using groundwater monitoring data (violet).

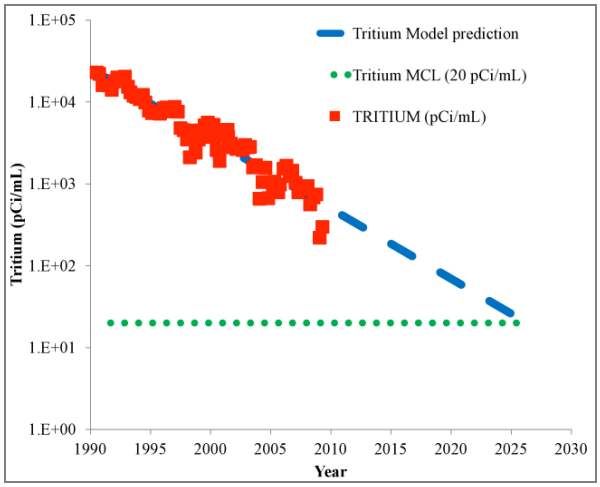

Figure 4: Exponential decay model applied to H-3 concentration monitored in well FSB 78.

5 NEXT STEPS TO IDENTIFY DATA OUTLIERS AND FLAGGING

Data outliers are measurements that are extremely large or small relative to the rest of the data and, therefore, are suspected of misrepresenting the population from which they were collected. The developed data management technology follows recommendations of the EPA documents EPA QA/G-9R[9] and EPA QA/G-9S[10] as well as EPA Data Qualifiers/flags for inorganics--http://www.caslab.com/EPA-Data-Qualifiers/, and for organics--http://www.caslab.com/EPA-Data-Qualifiers/

Small values below the MDL (minimum detection limit) and below the PQL(practical quantitation limit) are each flagged. Analyte concentrations that exceed the MDL but do not

exceed the PQL are often reported as estimates, and both the MDL and the PQL are a laboratory-specific numbers that change with time. The data management system also includes a comparison of concentrations of major contaminants in groundwater with the Maximum Contaminant Levels (MCLs) in drinking water, according to the EPA regulations (http://water.epa.gov/drink/contaminants/index.cfm#1).

In the next releases of the data management system, we plan to develop a basic statistical evaluation of environmental data, according to Chapter 4 of EPA QA/G-9S.

6 CONCLUSIONS

As part of the ASCEM project, we developed a new methodology for a subsurface data management system, including a new design of the database infrastructure. The resulting unified database system can represent data from heterogeneous sources having different formats. This approach not only allows the data to be stored in a uniform format, but also permits the automatic generation of views that drive the same kind of interfaces for all relevant data types. This approach makes it easy for users to find data of interest, use filters to select subsets of interest, and browse the data by visualizing time series plots and plume contour maps. When a user is satisfied with a selection, he/she can display the data in a tabular form and save the tabular data to be used in subsequent tasks. An example of using the developed data management system to support a case study is provided in this paper.

Acknowlegement. This work was supported by the U.S. Department of Energy EM-12 Office of Soil and Groundwater Remediation, ASCEM project, under Contract No. DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory.

REFERENCES

[1] Hanford site URL - http://www.hanford.gov
[2] Savannah River site URL - http://www.srs.gov/
[3] M. A. Phifer, M. R. Millings and G. P. Flach, “Hydraulic Property Data Package for the E-Area and Z-Area Soils , Cementitious Materials , and Waste Zones”, Washington Savannah River Company, Savannah River Site, (2006).
[4] A.D. Smits, M.K. Harris, K.L. Hawkins and G.P. Flach, “Integrated Hydrogeological Model of the General Separations Area (U); Volume 1: Hydrogeologic Framework (U)”. WSRC-TR-96-0399, Rev. 0., (1997).
[5] G.P. Flach and M.K. Harris, “ Integrated Hydrogeological Model of the General Separations Area (U); Volume 2: Groundwater Flow Model *(U)*”. WSRC-TR-96-0399, Rev. 1., (1999).
[6] D. Agarwal, M. Humphrey, N. Beekwilder, K. Jackson, M. Goode, and C. van Ingen. "A data centered collaboration portal to support global carbon-flux analysis." Concurrency and Computation: Practice and Experience - Successes in Furthering Scientific Discovery, Vol 22, Issue 17, pp 2323-2334, December 2010. doi: 10.1002/cpe.1600.
[7] M. Humphrey, D. Agarwal, and C. van Ingen. Fluxdata.org: Publication and Curation of Shared Scientific Climate and Earth Sciences Data. In Proceedings of the 5th IEEE international Conference on e-Science (e-science 2009). Dec 7-9, 2009. Oxford, UK.
[8] M. Humphrey, D. Agarwal, and C. van Ingen, "Publication and Curation of Large-Scale Shared Environmental Scientific Data," Microsoft Technical Report, MSR-TR-2008-93, August 2008.
[9] Data Quality Assessment: A Reviewer’s Guide, EPA QA/G-9R, EPA/240/B-06/002, 2006.
[10] Data Quality Assessment: Statistical Methods for Practitioners, EPA QA/G-9S, EPA/240/B-06/003, 2006