WM2013 Conference, February 24 – 28, 2013, Phoenix, Arizona USA

The Integrated Cloud-based Environmental Data Management System at Los Alamos National Laboratory – 13391

Karen Schultz Paige, Penny Gomez, Nita P. Patel, Chris EchoHawk, and Alison M. Dorries
Los Alamos National Laboratory, MS M996, Los Alamos, NM, 87544, ksp@lanl.gov

ABSTRACT

In today's world, instant access to information is taken for granted. The national labs are no exception; our data users expect immediate access to their data. Los Alamos National Laboratory (LANL) has collected over ten million records, and the data needs to be accessible to scientists as well as the public. The data span a wide range of media, analytes, time periods, formats, and quality and have traditionally existed in scattered databases, making comprehensive work with the data impossible.

Recently, LANL has successfully integrated all their environmental data into a single, cloud-based, web-accessible data management system. The system combines data transparency to the public with immediate access required by the technical staff. The use of automatic electronic data validation has been critical to immediate data access while saving millions of dollars and increasing data consistency and quality. The system includes a Google Maps based GIS tool that is simple enough for people to locate potentially contaminated sites near their home or workplace, and complex enough to allow scientists to plot and trend their data at the surface and at depth as well as over time. A variety of formatted reports can be run at any desired frequency to report the most current data available in the data base. The advanced user can also run free form queries of the data base. This data management system has saved LANL time and money, an increasingly important accomplishment during periods of budget cuts with increasing demand for immediate electronic services.

INTRODUCTION

The LANL site has been active since 1943 and data have been collected from the surrounding environment since the 1970s. LANL has conducted environmental cleanup operations on site since the early 1990s. The result of this has been the generation of more than ten million analytical records associated with samples collected on and off site to evaluate the environment surrounding LANL. These records cover a wide range of media including air, soil, sediment, biota, and water. Traditionally, the data was stored in small databases or in notebooks, binders, or reports and varied in terms of format and quality. Most researchers were not familiar with the data in databases other than their own, which made integrated use of that data almost impossible.

This need for integrated data was highlighted following the Cerro Grande forest fire which burned a significant area around LANL in 2000. An integrated environmental dataset was necessary to conduct a site risk assessment. Subsequently, LANL was required to make environmental data available to the public so that they could see what was present in the soil, air, and water around LANL and the surrounding communities. This resulted in a database called RACER that received copies of the datasets present in three large databases: the ambient air

database, the water quality database, and the soil, sediment, and biota database. This database was made available to the public to increase transparency of LANL's data process and increase the public's awareness of environmental sampling data.

In 2010, a Six Sigma process improvement study was completed on the entire data flow process, from sample collection, through submission to the analytical lab, receipt of the electronic data and loading it into the database, to the final copy of data delivered to a LANL staff member and copied to the RACER database. This study found several obstacles in the data flow process. The data flow process in use at the time required many weeks before scientists could look at their data. Sometimes there were delays due to errors in the Electronic Data Deliverable (EDD), the electronic record of the data from the analytical lab. In addition, each record was checked for quality using manual validation, which takes time to do correctly. The public was not easily able to access the data from LANL's environmental program, which interferes with the LANL's desire to be transparent to the public as much as possible. These issues and others frustrated the users of the environmental data system. The Six Sigma process identified several improvements that could be made, the most significant being the transition of the current database to a cloud-based data system, which would allow another significant improvement, autovalidation of data, to be implemented simultaneously.

DESCRIPTION

The requirements for a cloud-based environmental data system included:

- a publicly available view of the database that required no feeds and no transformations of the data,
- automatic electronic validation of all data upon upload,
- reduction in cost with an increase of uptime,
- free form querying, so that scientists could use their data fully,
- free design of new reports,
- accommodation of new data providers,
- the ability to plan data in the system, even planning for missing samples from a previous sampling campaign,
- graphical display of data on maps,
- sample tracking from planning through collection, submission, and upload,
- and invoice improvements that match costs to sample methods.

The cloud-based environmental database began with a competitively bid contract which was awarded to a small company, Locus Technologies. Security of the cloud-based system is ensured through the use of a Tier IV data center, which is the highest level of security and accessibility available. The data center is protected against power and communication failures and environmental and cybersecurity attack; redundant backups ensure seamless data protection and the highest level of uptime. The cloud-based system is web based and platform independent which means that all users, public and LANL internal, can access up-to-date LANL environmental data from any place at any time.

DISCUSSION

The cloud-based data system, Environmental Information Management or EIM, was initiated in February 2012. Since that time eleven million records have been loaded, which includes 27,000 locations and 250,000 samples. The requirements for the system as described above have been met, along with additional desired features.

The public view of the database is called Intellus and is available on the web at www.intellusnmdata.com . This is a nightly copy of the EIM database to ensure that the public has access to the most recent data, and that it is the exact same data that scientists are using to make cleanup decisions. For data that moves through the entire data process but is not environmental monitoring data, such as industrial hygiene data associated with facilities and personnel, or research data, a Non-Environmental site is available. For data collected for third parties, such as local or tribal entities, the data can be held for a review period before it is released to the public. Over a thousand unique users have accessed the public database since the inception of the database.

Although LANL's environmental data is made available to the public, LANL adheres to FISMA, the Federal Information Security Management Act, to ensure the security of data is not compromised. LANL complies with established security requirements for the protection and control of information and information systems including annual reviews of information security programs in keeping with FISMA. Because LANL's environmental data is available to the public through the Intellus website, it is not duplicated on the website www.data.gov, which is a repository of information made available to increase public access to data generated by the Federal Government.

One benefit of a cloud-based database is the potential to use automatic electronic data validation (data quality assessment). In the past, manual data validation by a team of chemists could cost up to two million dollars annually and could take up to four weeks to accomplish during the busy summer sampling period. This led to frustration on the part of data users who wanted to see their data as quickly as possible. The cloud based database provider was already providing automatic electronic data validation services to its other clients so it was very convenient, easy, and cost efficient to take advantage of their existing algorithms. The change to autovalidation saved time and money and improved consistency since it was no longer a manual task.

There are significant savings available in a cloud-based data system due to the aggregating of the costs of servers and IT staff over several clients. The cost savings associated with the implementation of EIM have been significant. Since implementation last year, one full time employee has been redeployed to new work. Savings on database development, computer support and maintenance, and manual validation is estimated at three million dollars annually, based on sampling levels of approximately 50,000 samples shipped annually. One contract for manual validation has been eliminated. In addition, delays in access to data has been reduced to less than one day, since data can be loaded directly into the system by the analytical lab themselves, and autovalidation occurs immediately upon loading, which gives power users almost immediate access to their data. Lastly, a cloud-based database has provided improvement

in uptime compared to a local server based system used in the past, giving clients more reliable access to their data at lower cost.

Free form querying is supplied through the use of an ad-hoc query tool. This tool allows users internal to LANL access to every field in the database for querying purposes. The Intellus database has many of the same tools as EIM, with the exception of the ad-hoc query tool. Approximately twenty to thirty power users use the internal data base (EIM) daily. The ad-hoc query tool is an adapted Structured Query Language (SQL) query tool, which uses the power of SQL statements without requiring the user to know SQL. This tool allows power users to make any kind of individualized query including costs, schedules, analytical results, geophysical parameters, field measurements, quality metrics, and more. The ad-hoc query tool provides users with the full power of the eleven million records loaded in the database thus far.

The database has the option to freely design queries and then make them available to other users. The ad-hoc query tool can be used to design the query. The query can then be saved so that users of varying permissions can view the query and run it. Alternatively, the query can be made available to the public for use. This is a particularly useful feature for datasets or data queries that may be complicated to replicate, or have a complicated set of fields, such as a query for forest fire data in an area that has samples over a range of media which may or may not have been affected by the fire.

One benefit of moving to a cloud based system is the flexibility for expansion of the system. With cloud-based servers, there is no limit on space because the servers are distributed physically throughout the country and more servers can be purchased and supported by the cloud provider, Locus Technologies in this case, as customer needs expand. In fact, additional data providers can benefit from the database structure that LANL has put in place in EIM and take advantage of the improvements that continue to be made in the system, thus saving local data providers money and time. The New Mexico Environment Department – Oversight Bureau has already loaded their environmental data into EIM and there is the possibility of other neighboring city, state, or tribal facilities entering their data into the data base if desired, while maintaining data control and access themselves.

The cloud-based system takes into account very complicated sample plans, specifically the requirement that incomplete sample collection be accounted for in future sample planning. It is extremely important that sample planning be done carefully for any sampling campaign. Good planning reduces errors by making sample collection and analysis requirements clear. It also makes invoicing more accurate. Environmental sample planning can be difficult because environmental samples are sometimes volume limited. Biota samples may be limited based on the body weight of an animal or limited populations. Water samples in dry climates like New Mexico are often volume limited due to weather. The topsoil in New Mexico can also be very shallow due to the rocky local geography. The sample planning system in EIM is sophisticated enough that any methods that could not be completed in a given sampling round are put forward in the next round until there is sufficient sample to complete that method, and the sampling can be fulfilled.

An appealing aspect of the cloud-based database is the GIS mapping tool available in EIM. The mapping tool uses Google map features that are generally familiar to the public. The mapping tool provides the user with a visual description of their data. Data can be plotted on the map and from the map can be charted and exported or organized into a downloadable table. An example of these features is shown in Figure 1 and Figure 2.

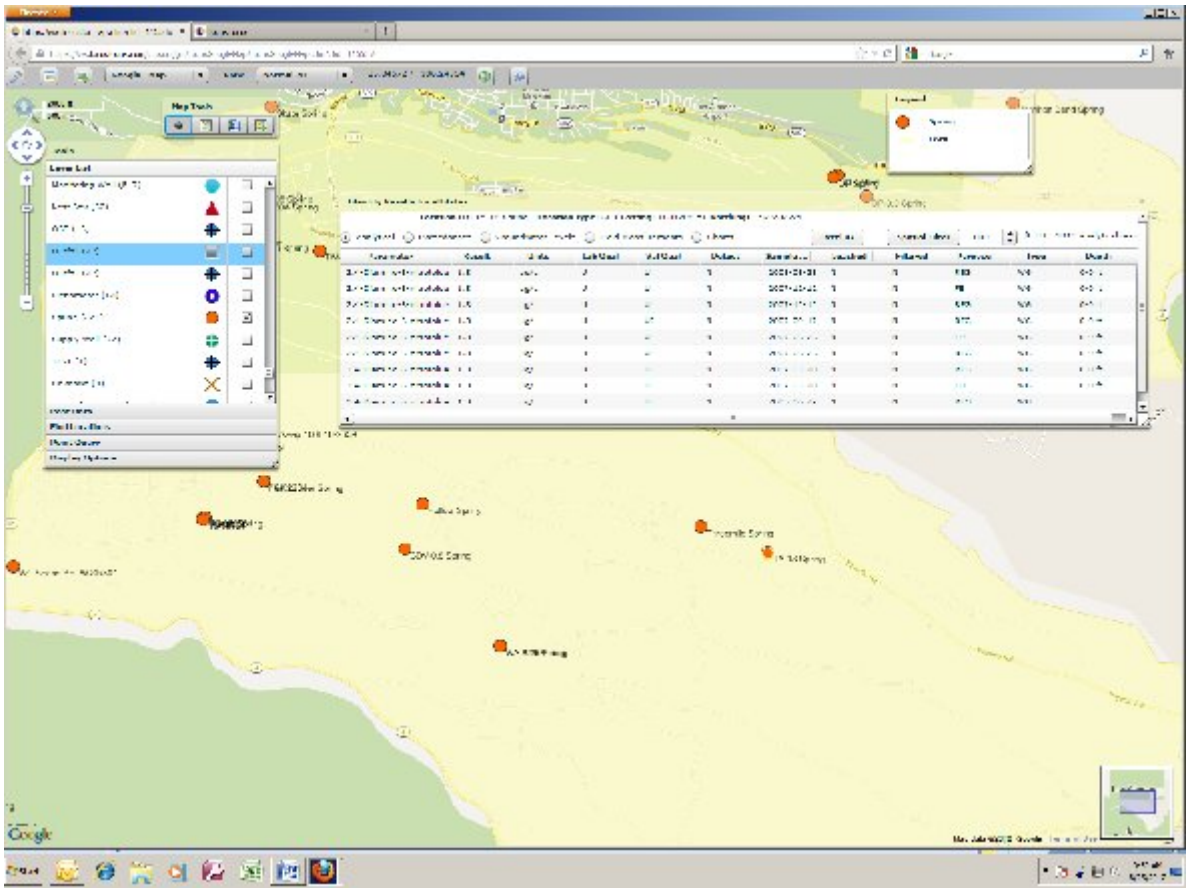

Fig. 1 Example of map option for tabulating data for locations shown on map.

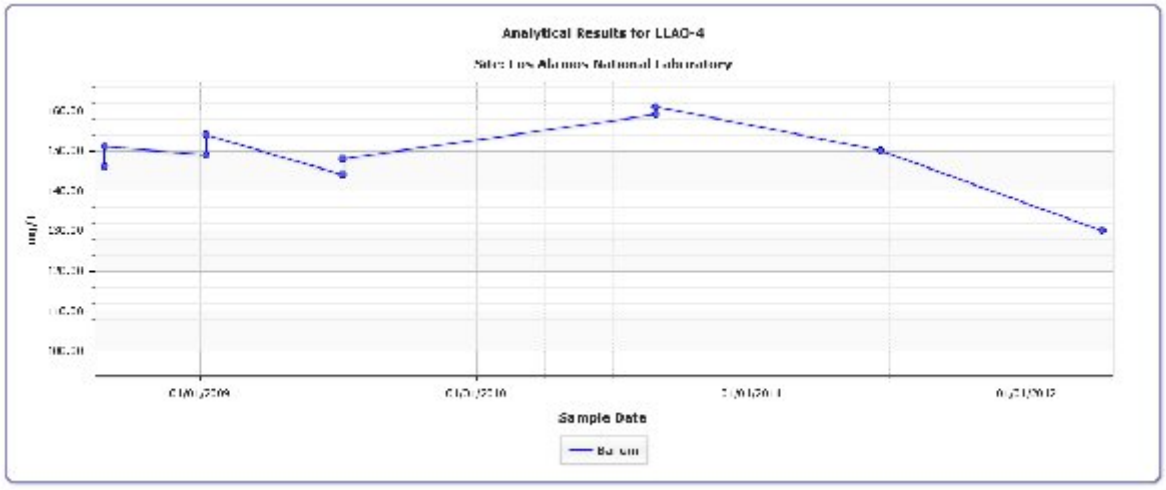

Fig.2 Example of exportable chart from EIM.
Another map option is to click on a point on a map and obtain environmental data near that point, which can help satisfy public concerns about contamination in and around homes and schools. Figure 3 shows an example of this option.

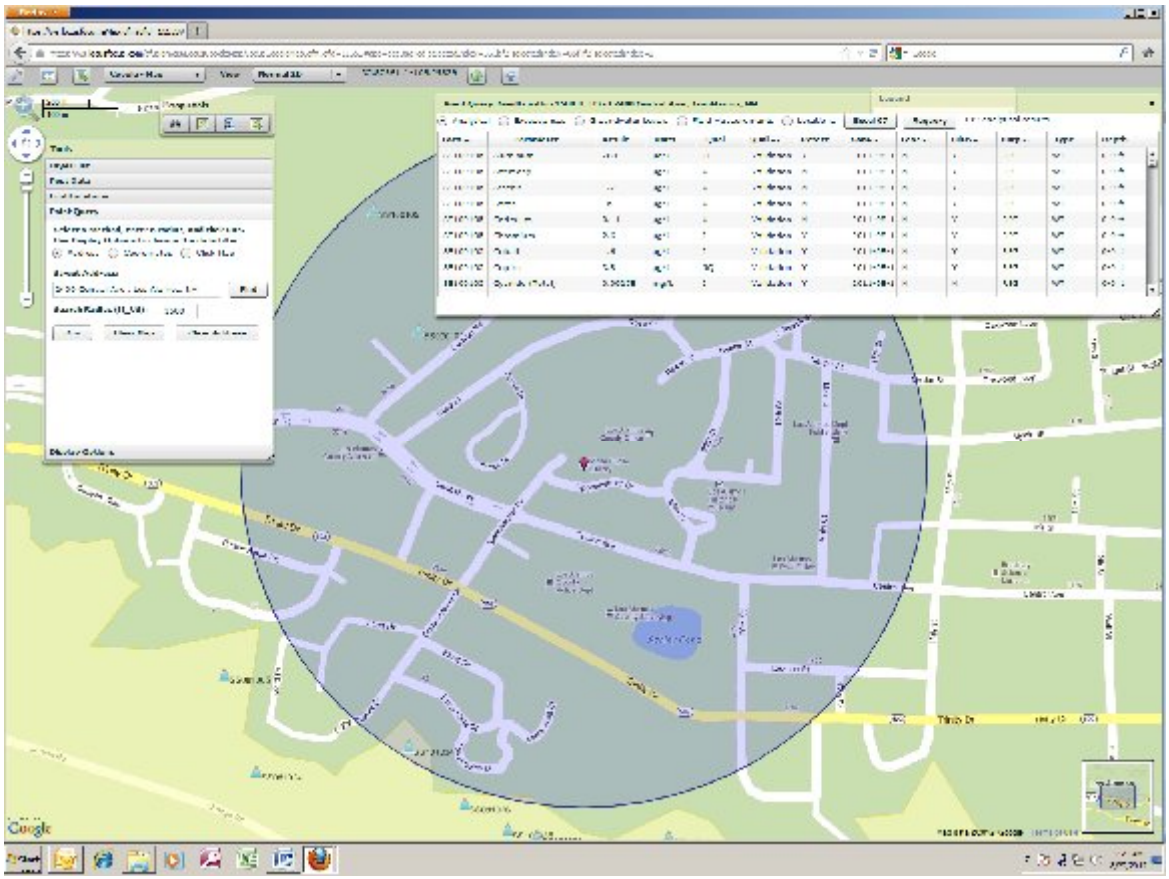

Fig. 3 Example of map option to post data near a point selected on the map.

Given the large volumes of samples moving through our environmental data system, keeping track of samples can be difficult. The EIM system has a complete tracking system that is transparent to the client as well as the Sample Management Organization staff. Field collection personnel enter the dates for sample collection and shipping, analysis date is automatically loaded by the analytical lab, and all remaining dates are populated by the database as the data moves through the database. It is possible for data owners to track their data on a daily basis from collection through shipping and analysis, to loading, validation, and final disposition.

As a result of the sophisticated sample planning process and the robust sample tracking system, the sample invoicing system in EIM can more accurately project sampling costs and identify invoicing and costing errors. Analytical costs within EIM are used following sample planning to provide project managers with an accurate estimate of the sample campaign costs as planned. Once a sample campaign has been completed and work has been billed, invoices can be accurately assessed based on the samples that have been actually received and the actual methods which were performed on those samples. These costs may differ from those estimated at planning but they accurately reflect what LANL must pay for the analytical services rendered. The invoicing system also provides a method to break out analytical costs on each invoice into

the projects that have incurred them.

LANL placed many requirements on the supplier of the EIM cloud-based database. The requirements for the system have been successfully met and improvements compared to past databases are already in place and being used to reduce cost, improve data access and quality, and increase functionality.

CONCLUSIONS

Overall, the EIM system has met our requirements and offered additional features as well. All media from three different environmental databases have been combined for the first time into one publicly accessible database which also meets the querying and reporting needs of technical power users. The savings from implementing the new system have been immediate and significant. The implementation of automatic data validation has improved access to final data by weeks, improved data quality and consistency, and saved up to two million dollars annually on validation costs. Eventually it will be possible for regulators and the public to generate reports from EIM without any paper copy, using only the most recent data, on any desired timescale or frequency.

[Los Alamos National Laboratory Unclassified Release Number LA-UR-12-26235]