LA-UR-11-10620
Approved for public release; distribution is unlimited.

Title:	Final Draft of RACER Audit
Author(s):	Paige, Karen Schultz Gomez, Penelope E.
Intended for:	Public RACER Audit Presentation, 2011-06-14 (Santa Fe, New Mexico, United States)

RACER Audit Response March 2011

This document describes the approach Waste and Environmental Services – Environmental Data and Analysis plans to take to resolve the issues presented in a recent audit of the WES-EDA Environmental Database relative to the RACER database.

A majority of the issues discovered in the audit will be resolved in May 2011 when the WES-EDA Environmental Database, along with other LANL databases, are integrated and moved to a new vendor providing an Environmental Information Management (EIM) system that allows reporting capabilities for all users directly from the database. The EIM system will reside in a publicly accessible LANL cloud-based software system. When this transition occurs, the data quality, completeness, and access will change significantly. In the remainder of this document, this new structure will be referred to as the LANL Cloud System

In general, our plan is to address the issues brought up in this audit in three ways: 1. Data quality issues such as units and detection status, which impinge upon data usability, will be resolved as soon possible so that data quality is maintained. 2. Issues requiring data cleanup, such as look up tables, legacy data, locations, codes, and significant data discrepancies, will be addressed as resources permit. 3. Issues associated with data feed problems will be eliminated by the LANL Cloud System, because there will be no data feed. As discussed in the paragraph above, in the future the data will reside in a publicly accessible system. Note that report writers may choose to convert, adapt, or simplify the information they receive officially through our data base, thereby introducing data discrepancies between the data base and the public report. It is not always possible to incorporate and/or correct these errors when they occur.

Issues in the audit will be discussed in the order in which they are presented in the audit report. Clarifications will also be noted as the audit report was a draft document, at the time of this response.

Background

In this section, one clarification should be noted. There is no requirement that RACER be the database for all environmental monitoring data. There are several examples of environmental monitoring data that have not been transferred to RACER, such a non-numeric field data and older data that lacks a pedigree or proper documentation.

Audit Objectives

One clarification should be made regarding the audit objective of ensuring data completeness in RACER. Prior to the audit, it was known to the audit team that the data in RACER were not complete. There are several data sets that have never been placed into RACER for a variety of reasons, such as some non-numeric field screening data. In addition, for some older legacy data no path exists to make the data more complete than it currently is, due to its age. Data completeness for all records, therefore, is an impossible goal for RACER to meet. However, that does not invalidate the goal of ensuring data transparency and ongoing data quality improvement.

Key Findings

The second finding indicates a possible misunderstanding during the audit. The finding that the audit team was unable to get a complete data set of measurement data is perhaps overly exaggerated. There was a disagreement with obtaining a data set for the ambient air monitoring program. LANL offered a complete dataset and it was refused. It was agreed a filtered data set would be delivered later. We believe that we satisfied the spirit of the audit data request.

Additionally, providing the fields to be queried during the audit prior to the audit is not sufficient to ensure that all data requested would be provided in the eight hour time frame of the audit. The queries done were large and complicated. We provided our best database programmer to be at our disposal during the entire eight hour audit. We do not feel a finding for data availability issues is appropriate.

Summary of Findings

Completeness of data in RACER

That data was found in RACER but not in the LANL database and that data were found in the LANL database but not in RACER was known to all before the audit was initiated. It was known that incorrect data for several samples associated with the Soil, Foodstuffs, and Biota (SFB) were loaded into RACER initially for system testing purposes and never removed. Consequently, the correct SFB data associated with those samples could not be loaded into RACER. However, the correct data will be loaded by April 30, 2011 and the incorrect SFB data deleted. It is valuable to know the general magnitude of discrepant data in the two databases but the existence of such a discrepancy is not a surprise.

Accuracy of Data in RACER

The issue of data accuracy in RACER will be resolved when the LANL Cloud System is put in place because the two listed reasons for discrepancies: a problem in the RACER update procedure (previously described) and changes made by the user but not propagated to RACER (explained in the "Completeness of Data in RACER" section below), will disappear.

Verification of Data Tables in RACER

Issues involving crosswalks will be resolved when the LANL Cloud System is in place, since there will be no need to crosswalk items. We appreciate having issues involving unit conversion brought to our attention as those items should be addressed promptly.

Verification of Detection Status and Data Usage

The logic for determining detection status and data usage will be reviewed prior to moving all data to the LANL Cloud System. Any errors or discrepancies will be resolved. Review of this logic is currently underway.

Verification of Supporting Information

The logic for determining hydrologic zones will be reviewed prior to moving all data to the LANL Cloud System. Any errors or discrepancies will be resolved. Review of this logic is currently underway.

Spot-checks of Public Reports

This section highlights some discrepancies that are within our control and some that are beyond our control. Those that can be addressed will be resolved as resources allow.

Preparation

LANL

The recommendations put forth in this section to review the feed script and review documentation will not need to be addressed since the LANL Cloud System will be in place before these changes can be made and the new system will negate the need for these items relating to the data feed or the RACER stand-alone data base.

Data Evaluation

Completeness of Data in RACER

We appreciate documentation of the fact that all the data in our database is not present in RACER. It has been known that some data are present in RACER but are not present in our database because they were incorrectly loaded. We know that some data is not going to RACER such as non-numeric field information and older data of questionable pedigree. In addition, some new data is not going to RACER directly while the Cloud System is being implemented, but will be manually loaded quarterly in the interim. When the LANL Cloud System is in place, all our data will be available to the public so the issue of completeness will not be a problem

Data found in RACER but not in LANL audit data set:
We have investigated the 448,814 records mentioned in this section in Table 1 as being in RACER but not in the LANL audit data set. Of that total, 23,084 records are the SFB data which will be purged and replaced in order to correct the SFB data originally loaded for test purposes (noted above under the "Completeness of Data in RACER" section). Another 363,322 records that are in LANL databases were excluded because preliminary data were excluded from the LANL audit set. However, some preliminary data is sent to RACER; for example, field data or data sampled at Los Alamos National Laboratory-EES6. These 363,222 records are both in RACER and LANL databases and are not of concern. 5,364 records had previously been flagged as ineligible for transfer after they had already been transferred to RACER. There is currently no process to inform the RACER database to remove these data once they are transferred to RACER which resulted in discrepancy between RACER and the LANL audit data set. 50,139 records had previously been deleted from LANL's internal databases due to legacy cleanup of duplicate records. These records will be flagged for removal from the RACER database. 108 records were assigned a code that means they were transferred to RACER but the code interfered with the process of generating the audit data set the audit data set was pulled. Lastly, 6,797 records are unaccounted for and their absence from the LANL audit data set cannot be determined at this time. It is not clear if they

will be evaluated since their absence cannot be explained. We speculate that a "process" error during the audit occurred and inadvertently missed compiling the records. However, we have directly accounted for 442,017 records claimed to be missing resulting in less than 2% of records remaining to be accounted for.

Data found in LANL audit data set but not RACER:

We have also investigated the 558,450 records that are listed in this section in Table 1 as being part of the LANL audit data set but not found in RACER. Of this total, 53,851 records were missing from RACER due to a "window" issue. The audit turned up an error in the manner of filtering and selecting data for transmission to RACER. A 30-day window was set for moving modified data to RACER and for deleting samples without results and results without samples. The missing data identified in the LANL audit data set were outside that window and therefore did not get transferred to RACER. The window has since been expanded up to a 365-day period to solve the problem going forward and the missing records have been uploaded to RACER. There were 24 results for ambient air data that had not yet gone to RACER, 28 results that were not eligible for migration and 58 that are in the water quality data base and would be migrated at the next load. The remaining 504,489 are duplicates which should not be in RACER. This explanation accounts for all records the audit team identified as data existing in the LANL databases but not in the RACER database.

LANL Environmental Measurement Data Findings

In summary, the issues involving missing or additional data in RACER and in our data base have been resolved to the best of our ability through resolution of the issues addressed above.

Summary of Findings

Several recommendations were made in this section. The first recommendation called for removal of records from RACER. In the past we have recommended a flush of RACER with a clean reload of data. This was never approved. As a policy, we do not remove any records from our data base, preferring instead to flag the data or move it to inactive tables. Records may be removed from RACER though. The remaining recommendations are not necessary in that the new LANL Cloud System will obviate keeping track of missing data.

Accuracy of Data

The issue of data accuracy in RACER will be resolved when the LANL Cloud System is put in place because there will be no discrepancies in data since all interested parties will be examining the same database.

Evaluation of Non-Transformed Data

The fields in the data base may have to be changed when the LANL Cloud System is put in place since many different groups will be using the data. This issue will have to be resolved during the transition.

Evaluation of Transformed Data

The fields in the data base that must be transformed may have to be changed when the LANL Cloud System is put in place since many different groups will be using the data. This issue will have to be resolved during the transition. At that time, we will reexamine the logic used to make the transformations to ensure that they are correct.

Verification of Analyte Crosswalks

As mentioned above, issues involving crosswalks will be resolved with the establishment of the LANL Cloud System. Since analyte names will not be transformed anymore, there will be no need for a crosswalk.

Verification of Unit Conversion Tables

We appreciate the audit drawing attention to an issue of unit conversions that we were unaware of. This issue will need to be dealt with prior to the LANL Cloud System transition. We will resolve these issues promptly.

Evaluation of Detection Status

We also appreciate the attention that the audit team has brought to the potential issue of incorrect detection status determination. We acknowledge the importance of identifying the detection status of an analyte correctly. It appears that there are discrepancies in how we describe detection status logic and how the logic is implemented. We are in the process of evaluating this issue right now and will continue to evaluate the logic so that it can be correctly implemented on transition to the LANL Cloud System. Once the logic is correctly identified, it will be applied to the data and corrections made.

The last finding in this section requests that a source field be added to data so that detection status can be more easily evaluated. This source flag already exists in the LANL data base and will be publicly accessible when the LANL Cloud System is in place.

Evaluation of Data Usage Flag

The issue of discrepancies in the data usage flag will be handled in a similar manner to the issue of detection status. The audit highlighted several potential inconsistencies in the data usage flag when compared to the logic that should be used to determine this flag. This issue will have to be reexamined with the transition to the LANL Cloud System so that all users will know the usage status of any given record. The logic will be evaluated for correct implementation and the logic will then be applied to the data in the LANL Cloud System.

Verification of Supporting Information

The supporting information reviewed in the audit, the SUBJECT_SAMPLE table and the hydrologic zone assignments will be significantly altered when the LANL Cloud System is installed. Therefore, current plans are not to make any significant changes to these items in anticipation of improved ways to obtain this kind of information in the LANL Cloud System.

Spot Checks of Data in Public Reports

As stated earlier, it is impossible for us to guarantee that all data presented in reports are present in RACER because there are no restrictions on data use. After a client receives data from the data base, they may choose to convert, adapt, or simplify the data for the end user. It is not always possible for those changes to be completely captured in the database.

Recommendations

Specific recommendations have been addressed in the text above. The recommendations in this section call for review of data management procedures and policies which will be done in anticipation of the LANL Cloud System. Another recommendation calls for changes to the system for sending and updating data. The LANL Cloud System will eliminate the need to send data or maintain data feeds.

Conclusions

The conclusions section highlights some problems in the databases that were known at the time of the audit. The first conclusion correctly states that there is a break in getting current soils, sediments, foodstuffs, and biota data into RACER. The second conclusion correctly states that there are data missing from RACER such as legacy SFB data and more. The third conclusion correctly states that data should be removed from RACER because it was known to be incorrect. The fourth conclusion is correct in that several discrepancies have been found between RACER and our database. Some of these discrepancies will be examined as resources are available and some discrepancies will disappear when the LANL Cloud System is installed. The final conclusion is correct in highlighting discrepancies in certain flags with concrete application logic. The logic for these flags will be reexamined with the installation of the LANL Cloud System, since the logic will be applied to similar flags in that system.

Generally, the issues brought up in this audit will be handled in three ways: 1. Issues addressing data quality, such as sample usage codes, unit conversions, and detection status, will be resolved as soon possible so that data quality is maintained. 2. Issues requiring data cleanup, such as look up tables, legacy data, locations, codes, will be addressed as resources permit. 3. Issues associated with data feed problems will be eliminated by the LANL Cloud System, because there will be no data feed. Significant issues involving discrepancies between data in reports and data in RACER will be addressed.