

DATA WAREHOUSES IN SUPPORT OF NUCLEAR EXPLOSION MONITORING RESEARCH

Julio Aguilar-Chang and Michael L. Begnaud

Los Alamos National Laboratory

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract No. W-7405-ENG-36

ABSTRACT

For the past six years, the Ground-based Nuclear Explosion Monitoring Research & Engineering (GNEM R&E) program has created and maintained an Oracle data warehouse at Los Alamos National Laboratory (LANL) to store and manipulate global and regional seismic research data. The LANL regional data holdings have focused on Eastern Asia. This data warehouse supports LANL scientists in their daily research efforts towards improving event location, discrimination, and characterization. It is also instrumental in providing data for research in focused areas of interest in Eastern Asia.

This paper will present a brief overview of how the LANL seismic research data warehouse has been used by researchers to achieve their goals of improving nuclear monitoring capabilities.

LANL's GNEM R&E seismic data warehouse uses Oracle's relational database management system (RDBMS) technology to store collections of global and regional seismic data that have been collected over the last six years from various sources, such as open global bulletins, academic colleagues, private contacts, and contractor collaborations. Two of the most challenging tasks of maintaining a functional data warehouse are the constant population of the database tables with the most up-to-date seismic information, and the care needed to integrate all the data sources into a cohesive database. We have received, and continue to receive, seismic bulletins from open sources such as the International Data Center (IDC) Reviewed Event Bulletin (REB), United States Geological Survey (USGS) Earthquake Data Reports (EDR), and International Seismic Centre (ISC) event bulletins. The event location solutions reported in these bulletins, corresponding phase arrivals, and network and station magnitude information are reconciled into single events; thus, researchers are able to compare bulletin information as reported by different organizations for a given event.

The reconciled database tables assist researchers working in event relocations by making available to them a merged set of phase arrivals and ground-truth information from global and regional sources. The location effort has, in fact, created its own location data warehouse by combining all available phase arrival information and manipulating the database tables to create consistent arrivals and improve event relocations.

The data warehouse is also important for event identification, discrimination, and characterization work because it provides consistent location information needed in the seismic analysis code (SAC) headers in the waveform data file. We use in-house programs—primarily written in Perl—to query database tables for preferred event location solutions, and to imbed these solutions into the SAC headers. The event discrimination and characterization work makes extensive use of the SAC headers to correctly tie measurements to events. The ability to store amplitude measurements in the research data warehouse, and to use those measurements in MDAC (Magnitude and Distance Amplitude Corrections) and coda magnitude work, provide us the means to tie all these results to specific events. The source information and resulting measurements are kept in a consistent and integrated framework.

OBJECTIVE

This paper will present an overview of the LANL seismic research data warehouse, and how LANL researchers use it to help them calibrate seismic stations used by the United States government in their mission of monitoring for clandestine nuclear explosions.

RESEARCH ACCOMPLISHED

Introduction to the Data Warehouse

For the past six years, the Ground-based Nuclear Explosion Monitoring Research & Engineering (GNEM R&E) program has created and maintained an ORACLE data warehouse at LANL to store and manipulate global and regional seismic research data. The LANL regional data holdings have focused primarily in Eastern Asia. This data warehouse supports LANL scientists in their daily research efforts towards improving algorithms and seismic models for event location, discrimination, and characterization. The data warehouse maintained at LANL uses the same technology as the data warehouses at the two other National Nuclear Security Administration (NNSA) laboratories that are part of the GNEM R&E program, as well as at the Air Force Technical Applications Center (AFTAC), the ultimate customer of our research efforts. The three NNSA laboratories and AFTAC use ORACLE's relational database management system (RDBMS) technology. The laboratories have maintained close contact with AFTAC to ensure that we are using the same version of the database server, thus allowing for easy exchange of data. Likewise, the design of the database tables is uniform throughout the GNEM R&E program, which helps to ease collaboration between researchers and guarantees consistency in the presentation of the final results to the user. The current agreed upon format for database tables is documented in *National Nuclear Security Administration Knowledge Base Core Table Schema Document* [Carr].

Contents of the Data Warehouse

The LANL GNEM R&E seismic data warehouse stores global and regional seismic data that have been collected over the last six years from various sources such as open global bulletins, academic colleagues, private contacts, and contractor collaborations. Our data warehouse holds over 114 gigabytes of data and provides direct access to over 160,000 seismic waveforms. We have received, and continue to receive, seismic bulletins from open sources such as the International Data Center (IDC) Reviewed Event Bulletin (REB), United States Geological Survey (USGS) Earthquake Data Reports (EDR), and the International Seismic Centre (ISC). Over the years, the number of distinct event locations and phase arrivals available from the global bulletins has increased at an amazing rate. Figure 1 below shows the cumulative number of event origins collected from global and other bulletins per year. Similarly, Figure 2 shows the cumulative number of collected phase arrival picks from available catalogs per year, and the number of reporting stations.

Two of the most challenging tasks of maintaining a functional data warehouse are the constant population of the database tables with the most up-to-date seismic information and the care needed to integrate all the data sources into a cohesive database. We follow various steps of collecting and integrating seismic data into the LANL GNEM R&E data warehouse depending on the source of the data. We electronically collect weekly and monthly bulletin reports from the USGS EDR, and daily event bulletins from the IDC REB. Once a week and once a month we run a program to query the USGS EDR web site for new bulletins. Likewise, once a day we query the IDC data source for new events. The bulletins are sent to LANL through electronic mail and stored on a disk as ASCII flat files. We then run Perl programs, written at LANL, to parse the information from the native bulletin format and transform them into KB schema format flat files. The EDR and REB Perl parsers look at the contents of the database tables to determine if events already exist and will make the appropriate connection to match newly collected EDR and REB events to existing ones. This process also assigns the preferred location solution for an event through the use of an origin-location-author ranking list. This list was created at LANL to establish the rank of origin location authors when an event has more than one location contributed by different sources. This ranking list was created based on prior knowledge of the data catalogs that we have received, and on years of researcher experience in dealing with origin locations from diverse sources. The LANL EDR and REB Perl parsers create new *event* KB schema flat files with updated preferred origin authors, as well as *origin*, *origerr*, *assoc*, *arrival*, *netmag*, and *stamag* ASCII text files. The data in these files are merged with existing data in the database tables and are immediately made available to LANL researchers. We also use built-in ORACLE database server functionality to ensure quality control of these new data so that duplicate information is not inserted into the existing tables.

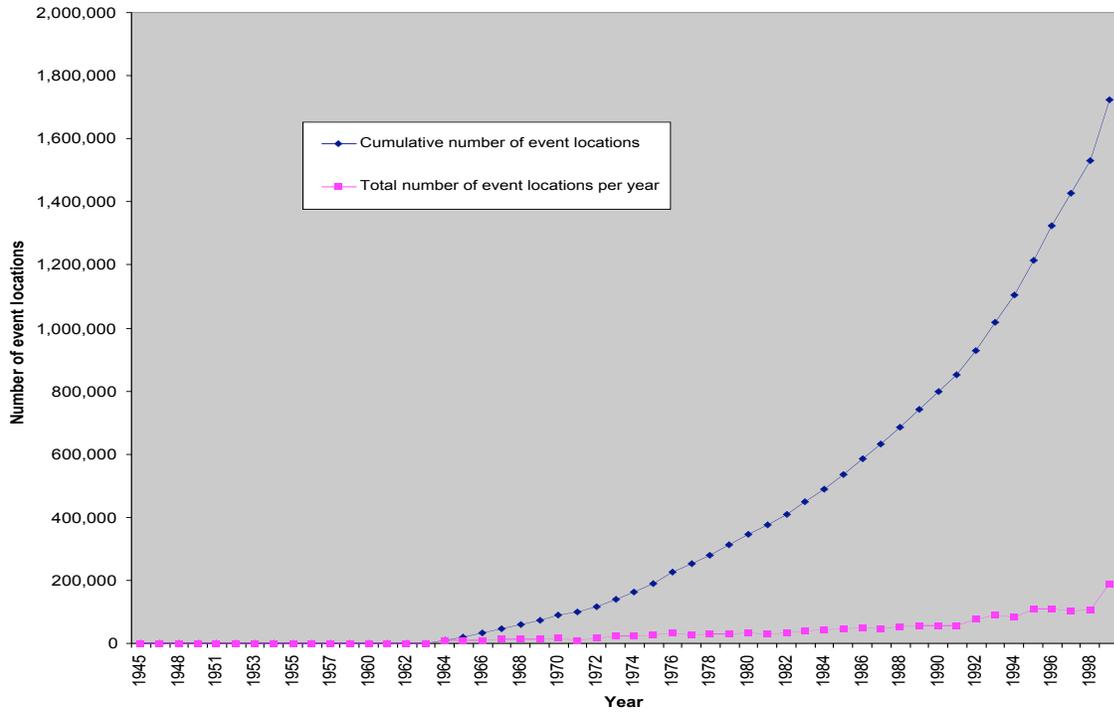


Figure 1. Total and cumulative number of reported event location solutions.

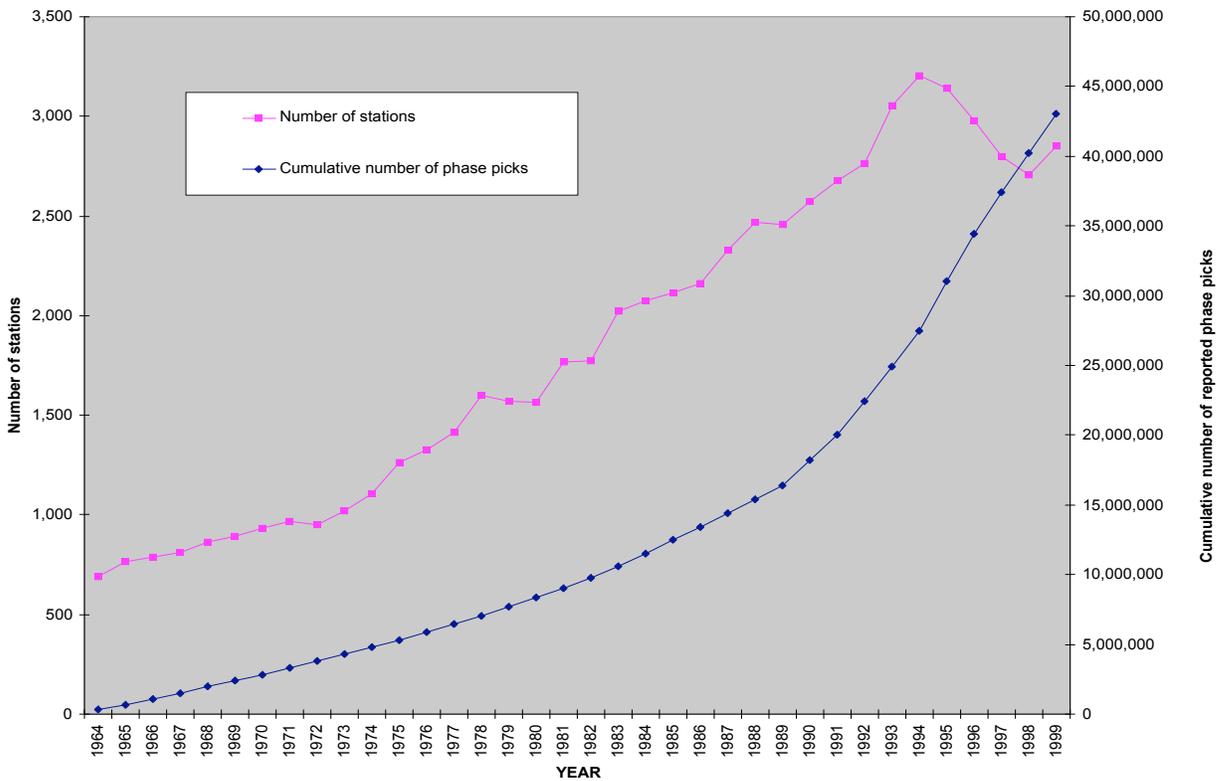


Figure 2. Cumulative number of stations and reported phase picks.

We follow a similar process to collect data from the ISC seismic bulletins and update our data warehouse. However, we use the program ORLOADER created at Lawrence Livermore National Laboratory LLNL to parse, load, and reconcile the information in the native ISC bulletin format with the existing data in the database tables. Similar to the parsing and loading of EDR and REB bulletins, the origin-author ranking table is also used by ORLOADER to update the preferred origin assigned to an existing event when loading ISC data.

Periodically we collect binary waveform files in seed format from organizations such as the Incorporated Research Institutions for Seismology (IRIS) Data Center. These data are transformed at LANL into SAC format waveforms. The data warehouse is used to populate the SAC headers with event, origin, and waveform identification information, which tie a particular binary waveform file to an event and preferred origin location in the data warehouse. Once the SAC headers have the correct identification information, our researchers make body-wave phase picks. These LANL-generated phase arrival data are merged and reconciled with existing data in the data warehouse and made available to the event location researchers to use in their relocation efforts. The SAC waveforms are also available for amplitude processing, which is used later on in event discrimination research. Figure 3 shows a graph depicting the total cumulative number of waveforms at LANL per year and the total cumulative number of phase picks made by our researchers. Figure 4 shows the cumulative number of raw amplitude measurements and amplitude corrections measured at LANL over the years for most of our waveform holdings.

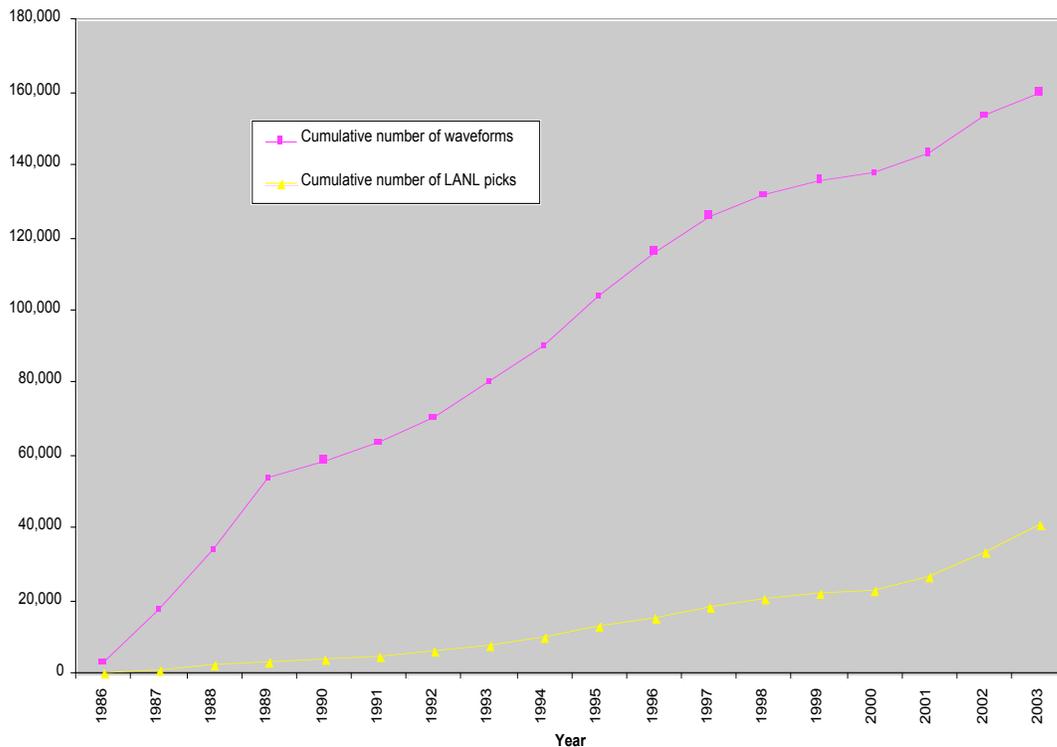


Figure 3. Cumulative number of waveforms (top line) and LANL picks (lower line).

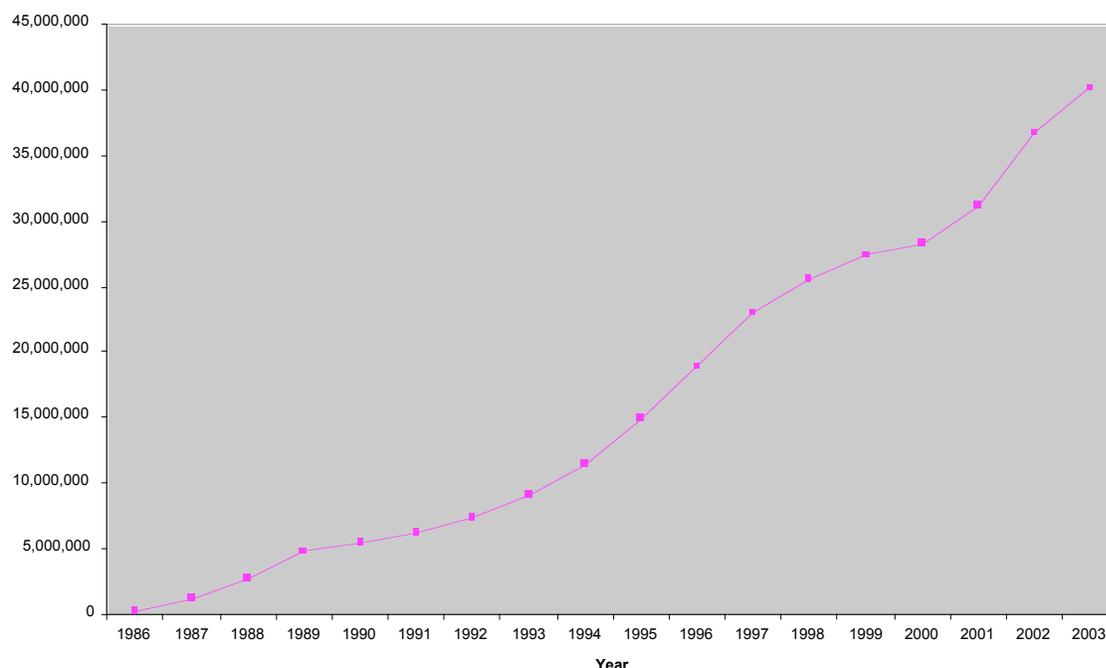


Figure 4. Cumulative number of amplitude measurements.

Using the Data Warehouse for Event Location Research

Merged global and regional phase arrival and ground-truth source data is available in our research data warehouse to assist researchers in their event location efforts. The location effort has, in fact, created its own database tables by combining all available phase arrival information and manipulating it to create consistent arrivals and improve event relocations that lead to better correction surfaces. Researchers use the reconciled database tables that contain global and regional bulletin information, as well as LANL-generated phase arrival and waveform data, to create new *arrival* and *assoc* tables specific for relocation purposes. When working with the location-specific database tables, the *assoc* table contains only the phase arrivals associated with the preferred origin location for a given event based on a pre-determined LANL-generated ranking table. One event can have one or more origin locations, and each origin location corresponds to a different author, i.e., the USGS EDR, IDC REB. This *assoc* table also contains all phase arrival picks made at LANL. Seismic phases in the *arrival* table are renamed so that all phases associated with a given event have unique descriptive names, which is a requirement of the EvLoc program used in location tasks. The renaming of the seismic phases follows a pre-determined ranking scheme, in which LANL picks are ranked highest and phase names remain unchanged. An example is shown in Table 1.

Table 1: Example of phase renaming

Phase pick author	Rank	Original phase			Renamed phase		
		time 1	time 2	time 3	time 1	time 2	time 3
LANL	1	Pn	Pg		P	Pg	
REB	2	Pg	P	Pn	Pg2	P2	Pn
EDR	3	Pn	P	Pg	Pn2	P3	Pg3

The phase renaming process is necessary to provide the best distinct phase arrival times to the programs used in location. By doing this process, we are compiling a database with the best distinct seismic phase arrivals, so that for a given event at a given station there is only one of each P, Pg, and Pn phase arrival.

Using the Data Warehouse for Event Discrimination Research

The event discrimination and magnitude research conducted at LANL benefits directly and indirectly from the data warehouse. The event discrimination work uses data stored in database tables as input for processing, while magnitude work uses the data contained in the headers of the SAC-format waveform files that are populated using data in the database. The final products of the event discrimination work are MDAC (Magnitude and Distance Amplitude Corrections) correction parameters for given event/station paths and amplitude corrections for raw amplitude measurements. One of the final products of the magnitude research work is the computation of network and station coda moment magnitudes.

As a first step, the event and station location information, magnitude estimates, and reporting catalog name contained in the headers of the SAC-format waveform files are compared against data in the database tables. When a match is found, the waveform headers are updated with the event, origin, and waveform identification numbers that tie a particular waveform to a specific event and location solution in the database. The researcher uses this header information, together with the waveform file, to make the calculations needed to estimate network and station coda moment magnitudes for a particular event. These results are inserted into the database tables' *netmag* and *stamag* that contain network and station event magnitudes respectively. At the same time, raw amplitude measurements are made on the waveforms for a variety of phases and frequency bands, and the results are then incorporated into the appropriate database tables. These amplitude measurements, as well as the magnitude estimates, are tied through a unique origin identification number to the corresponding event origin solution. At this point, the waveforms, phase arrivals, associated preferred event location, amplitude measurements, and magnitude estimates are all tied together through the event and origin identification numbers; thus, we are taking advantage of the relational nature of these data by storing them in a relational database server. We use programs written in Perl that query the database tables for event locations, magnitudes, and raw amplitude measurements, and the output is a set of ASCII flat files that are used in in-house written Matlab codes to perform the actual computations of MDAC parameters and amplitude corrections. The resulting MDAC amplitude corrections are then inserted back into the corresponding database tables and tied to existing raw amplitude measurements by using the event and origin identifiers.

Integration and Reconciliation of Data for Knowledge Base Deliverables

A large portion of the research results is stored in database tables in our data warehouse. New event location solutions from event relocation work are stored in KB schema *origin*, *origerr*, and *assoc* tables. The coda moment magnitudes computed for selected events are stored in KB schema *netmag* and *stamag* tables. The raw amplitude measurements and amplitude corrections are stored in custom tables *nnsa_amp_descript* and *nnsa_amplitude*. Finally, the results of MDAC research are stored in custom tables *mdac_fi* and *mdac_fd*. These research results are most useful to our customer as part of an integrated KB. An integrated KB gives the final user centralized access to all research results associated with the GNEM R&E program. The research results produced at LANL are integrated as part of a station-centric approach; this approach gives the user tremendous flexibility to view and access the data in various forms. Graphical user interface (GUI)-based tools have been developed to access these data, which makes the mining of the KB extremely efficient.

The database tables containing our research results are integrated and reconciled with those from other NNSA laboratories to accomplish the goals described above. LANL's research products are delivered to the customer through KB deliveries. There are two documents that describe in detail the procedures followed to compile and create a KB deliverable. These are the *Knowledge Base Contributor's Guide* [Carr], and *The Integration Process for Incorporating Nuclear Explosion Monitoring Research Results into the National Nuclear Security Administration Knowledge Base* [Gallegos, et. al]. The step-by-step description of the steps that the NNSA laboratories follow to meet the requirements for delivering research products to the KB is beyond the scope of this paper. The two documents mentioned above, as well as other papers in this conference, describe these steps in detail.

CONCLUSIONS AND RECOMMENDATIONS

The bulletins maintained in our warehouse report information such as event location solutions, phase arrival times, and network and station magnitudes. These data are reconciled into single events, and researchers are able to compare the information obtained from many bulletins as reported by different organizations for a given event, giving them the ability to quickly evaluate which information is most appropriate for the event. For example, for an

event a researcher may choose to use the location as reported from organization A and the network magnitude as reported from organization B. This choice is made possible by using the relational power of the data warehouse management system.

The reconciled database tables help researchers working on event location problems achieve improved event relocations by making available a merged set of global and regional phase arrivals and ground-truth information. The location effort has, in fact, created its own location database by combining all available phase arrival information.

The data warehouse is also instrumental in the event identification, discrimination, and characterization work because it provides consistent location information needed to populate the headers in the SAC-formatted waveform data files. The event discrimination and characterization work makes extensive use of the SAC headers to correctly tie measurements to events. The ability to store amplitude measurements in the data warehouse, and use those measurements in MDAC (Magnitude and Distance Amplitude Corrections) computations, provides researchers the means to tie all their results to specific events. The source information and resulting measurements are kept in a consistent and integrated framework.

ACKNOWLEDGEMENTS

The people who contribute to and maintain the Los Alamos National Laboratory research data warehouse are: Julio C. Aguilar-Chang, Michael L. Begnaud, Hans E. Hartse, Steven R. Taylor, W. Scott Phillips, George E. Randall, Richard J. Stead, Thomas L. Riggs, and Diane F. Baker.

REFERENCES

Carr, D. (09/02), *Knowledge Base Contributor's Guide*, Sandia National Laboratories report SAND2002-2771 (available at <https://www.nemre.nnsa.doe.gov/cgi-bin/prod/nemre/index.cgi?Page=Knowledge+Base>).

Carr, D. (09/02), *National Nuclear Security Administration Knowledge Base Core Table Schema Document*, Sandia National Laboratories report SAND2002-3055 (available at <https://www.nemre.nnsa.doe.gov/cgi-bin/prod/nemre/index.cgi?Page=Knowledge+Base>).

Gallegos, D. et. al. (09/02), *The Integration Process for Incorporating Nuclear Explosion Monitoring Research Results into the National Nuclear Security Administration Knowledge Base*, Sandia National Laboratories report SAND2002-2772 (available at <https://www.nemre.nnsa.doe.gov/cgi-bin/prod/nemre/index.cgi?Page=Knowledge+Base>).