LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Enhancing Seismic Calibration Research Through Software Automation

S.D. Ruppert, D.A. Dodge, A.B. Elliott,
M.D. Ganzberger, T.F. Hauk, E.M. Matzel, and
F. Ryall

July 2004

ENHANCING SEISMIC CALIBRATION RESEARCH THROUGH SOFTWARE AUTOMATION

Stanley D. Ruppert, Douglas A. Dodge,Annie B. Elliott, Michael D. Ganzberger, Teresa F. Hauk,
Eric M. Matzel, and Floriana Ryall

Lawrence Livermore National Laboratory

ABSTRACT

The National Nuclear Security Administration (NNSA) Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEM R&E) Program has made significant progress enhancing the process of deriving seismic calibrations and performing scientific integration with automation tools. We present an overview of our software automation efforts and framework to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research. The software and scientific automation initiatives directly support the rapid collection of raw and contextual seismic data used in research, provide efficient interfaces for researchers to measure/analyze data, and provide a framework for research dataset integration. The automation also improves the researcher's ability to assemble quality controlled research products for delivery into the NNSA Knowledge Base (KB). The software and scientific automation tasks provide the robust foundation upon which synergistic and efficient development of, GNEM R&E Program, seismic calibration research may be built.

The task of constructing many seismic calibration products is labor intensive and complex, hence expensive. However, aspects of calibration product construction are susceptible to automation and future economies. We are applying software and scientific automation to problems within two distinct phases or "tiers" of the seismic calibration process. The first tier involves initial collection of waveform and parameter (bulletin) data that comprise the "raw materials" from which signal travel-time and amplitude correction surfaces are derived and is highly suited for software automation. The second tier in seismic research content development activities include development of correction surfaces and other calibrations. This second tier is less susceptible to complete automation, as these activities require the judgment of scientists skilled in the interpretation of often highly unpredictable event observations. Even partial automation of this second tier, through development of prototype tools to extract observations and make many thousands of scientific measurements, has significantly increased the efficiency of the scientists who construct and validate integrated calibration surfaces. This achieved gain in efficiency and quality control is likely to continue and even accelerate through continued application of information science and scientific automation.

Data volume and calibration research requirements have increased by several orders of magnitude over the past decade. Whereas it was possible for individual researchers to download individual waveforms and make time-consuming measurements event by event in the past, with the Terabytes of data available today, a software automation framework must exist to efficiently populate and deliver quality data to the researcher. This framework must also simultaneously provide the researcher with robust measurement and analysis tools that can handle and extract groups of events effectively and isolate the researcher from the now onerous task of database management and metadata collection necessary for validation and error analysis. We have succeeded in automating many of the collection, parsing, reconciliation and extraction tasks, individually. Several software automation prototypes have been produced and have resulted in demonstrated gains in efficiency of producing scientific data products. Future software automation tasks will continue to leverage database and information management technologies in addressing additional scientific calibration research tasks.

OBJECTIVES

The National Nuclear Security Administration (NNSA) Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEM R&E) Program has made significant progress enhancing the process of deriving seismic calibrations and performing scientific integration with automation tools. We present an overview of our software automation efforts and framework to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research and the attributes required to construct next generation data acquisition. The scientific automation engineering and research will need to provide the robust hardware, software, and data infrastructure foundation for synergistic GNEM R&E Program calibration efforts. The current task of constructing many seismic calibration products is labor intensive and complex, hence expensive. However, aspects of calibration product construction are susceptible to automation and future economies. Data volume and calibration research requirements have increased by several orders of magnitude over the past decade. We have succeeded in automating many of the collection, parsing, reconciliation and extraction tasks, individually. Several software automation prototypes have been produced and have resulted in demonstrated gains in efficiency of producing scientific data products. In order to fully exploit voluminous real-time data sources and support new requirements for time critical modeling, simulation, and analysis, a more scalable and extensible computational framework will be required.

RESEARCH ACCOMPLISHED

The primary objective of the Scientific Automation Software Framework (SASF) efforts are to facilitate development of information products for the Ground-Based Nuclear Explosion Monitoring Research and Engineering (GNEM R&E) regionalization program. The SASF provides efficient access to, and organization of, large volumes of raw and derived parameters, while also providing the framework to store, organize, integrate and disseminate information products for delivery into the National Nuclear Security Administration Knowledge Base (NNSA KB). The current framework supports integration, synthesis, and validation of the various different information types / formats required by each of the seismic calibration technologies (Figure 1). For example, the seismic location technology requires parameter data (site locations, bulletins), time-series data (waveforms), and produces parameter measurements in the form of arrivals, gridded geo-spatially registered corrections surfaces and uncertainty surfaces. Various tools and information processing frameworks (relational database (RDB), Geographical Information System (GIS), and associated product/data visualization and data management tools (e.g. RBAP, KBALAP, KBCIT, DM).

These information management and scientific automation tools used together within specific seismic calibration processes to support production of tuning parameters for the United States Nuclear Seismic Monitoring System run by the Air Force (Figure 2). The calibration processes themselves appear linear (Figure 2) beginning with data acquisition extending through reconciliation, integration, measurement/simulation through to construction of calibration / run-time parameter products. Efficient production, however, of calibration products requires extensive synergy and synthesis between not only data-types (Figure 1), but also between measurements and results derived from the different calibration technologies (e.g. Location, Identification, Detection) (Figures 1 and 2). Even with successful implementation of automation within many of the individual steps, the current infrastructure will not scale to handle an order of magnitude additional data nor extend to handle time critical data acquisition or analysis. This lack of scalability and flexibility limits efficient production and delivery of run-time calibrations to the operational seismic monitoring pipeline (Figure B – Bottom) as a large manual effort is still required to acquire and integrate streaming (10 – 20 GB/day) signals with associated metadata. This synergy and synthesis between complex tools and very large datasets is critically dependent on having a scalable and extensible unifying framework. These requirements of handling large datasets in diverse formats and facilitating interaction and data exchange between tools supporting different calibration technologies led to an extensive scientific automation software engineering effort to develop an object oriented database centric framework (Figure 3) as an unifying foundation.

Figure 1. The Scientific Automation Software Framework provides a unifying framework for contextual/reference data and information products.

Figure 2: Summary of the processes of data collection, research and integration within the LLNL calibration process that result in contributions to the NNSA KB. The relationships of the current LLNL calibration tools, scientific automation tools, and database coordination framework to those involved in the assembly of the NNSA KB or within the AFTAC operational pipeline are delineated.

Scientific Automation Software Tools

Information products created using the LLNL SRDB may be grouped under two major categories or tiers: Tier 1 - primary data products and Tier 2 - derived products. In order to calibrate seismic monitoring stations, the LLNL SRDB must incorporate and organize the following categories of primary and derived measurements, data and metadata:

Tier 1: Contextual and Raw Data
Station Parameters and Instrument Responses
Global and Regional Earthquake Catalogs
Selected Calibration Events
Event Waveform Data
Geologic/Geophysical Data sets
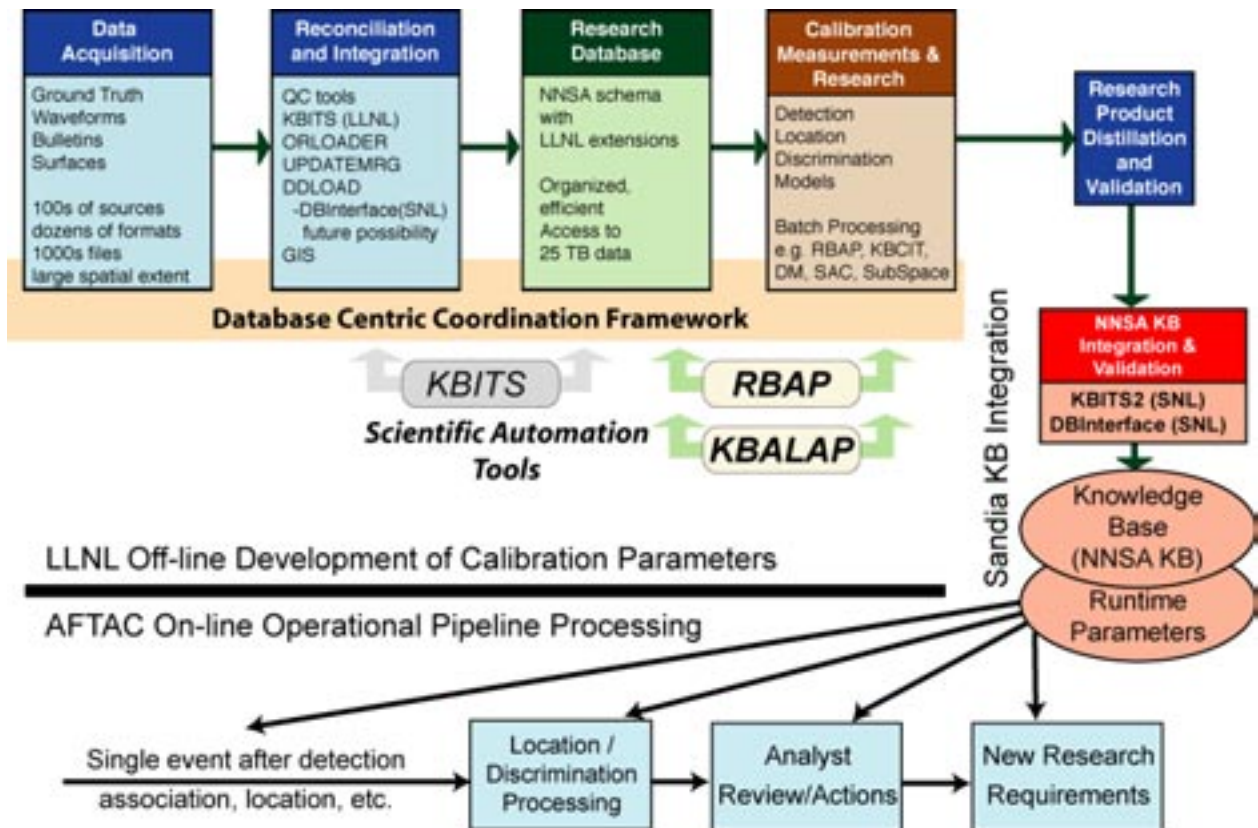Geophysical Background Model

Tier 2: Measurements and Research Results
Phase Picks
Travel-time and Velocity Models
Rayleigh and Love Surface Wave Group Velocity Measurements
Phase Amplitude Measurements and Magnitude Calibrations
Detection and Discrimination Parameters

Automating Tier 1

Corrections and parameters distilled from the calibration database provide needed contributions to the NNSA KB for the ME/NA/WE region and will improve capabilities for underground nuclear explosion monitoring. The contributions support critical functions in detection, location, feature extraction, discrimination, and analyst review. Figure 2 outlines the processes of data collection, research and integration within the LLNL calibration process that result in contributions to the NNSA KB and the relationship of the LLNL calibration tools to those involved in the assembly of the NNSA KB or within the AFTAC operational pipeline. Within the major process steps (Data Acquisition, Reconciliation/Integration, Calibration Research, Product Distillation) are many labor intensive and complex steps. The previous bottleneck in calibration process was in the reconciliation/Integration step (Figure 2). This bottleneck became acute in 1998 and the KBITS suite of automated parsing, reconciliation, and integration tools for both waveforms and bulletins (ORLOADER, DDLOAD, UpdateMrg) were developed. The KBITS suite provided the additional capability required to integrate data from many datasources and external collaborations. Data volumes grew from the 11,400 events / 1 million waveforms in 1998 to the 4 million events / 60 million waveforms today (e.g. Ruppert et al., 1998, O'Boyle et al. 2003).. This rapid increase in stored parameters soon led to two new bottlenecks hindering rapid development and delivery of calibration research

Automating Tier 2

As the number of data sources required for calibration increased in number and source location, it became clear that the manual and labor intensive process of humans transferring thousands of files and un-manageable metadata could not keep the KBITS software fed with data to integrate nor could the seismic research efficiently find, retrieve, validate, or analyze the raw parameters necessary to effectively produce seismic calibrations in an efficient manner. Significant software engineering and development efforts were applied to address this critical need to produce software aids for the seismic researcher. Two scientific automation tool prototypes (RBAP, KBALAP) (Figure 2) are under development for Seismic Location and Seismic Identification calibration tasks.

Both of these prototypes include methods and aids for efficiently extracting groups of events and waveforms from the millions contained in the SRDB and making large numbers of measurements with metadata in a batch mode. The concept of event sets (groups of related seismic events or parameters that can be processed together, e.g. either station centric or event centric) was introduced as previous SAC scripts and macros could not scale to the task.

The KBALAP Program

The Knowledge Base Automated Location Assessment and Prioritization (KBALAP) program is a set of database services and a client application that combine to efficiently produce location ground truth data that can be used in the production of travel time correction surfaces and as part of the preferred event parameters used by other tools in our processing framework.

The part of KBALAP that runs as a database service is responsible for evaluating bulletin and pick information as it enters the system to identify origin solutions that meet pre-defined GT criteria with no further processing, and to identify events that would likely meet a pre-defined GT level if a new origin solution was produced using available arrivals. The database service is also responsible for identifying events that should have a high priority for picking based on their existing arrival distribution, and the availability of waveform data for stations at critical azimuths and distances.

The interactive portion of KBALAP has three principal functions. These are
  • interactive production of GT origins through prioritized picking and location,
  • interactive specification of GT-levels for epicenter, depth, origin time, etype, and
  • batch-mode location of externally-produced GT information.
The first of these capabilities allows the user to view epicenters and GT information on a map based on selection criteria input by the user. The user can select any GT or potential GT event and observe the distribution of stations with picks and stations with available waveforms. The user can select any station with available waveforms and open a picker with any current picks displayed. There the user can adjust existing picks, add new picks, mark bulletin picks as unusable, and relocate the event. A new GT level is calculated, and the user can choose to accept that origin solution and GT level or continue working with other stations.

The interactive GT entry mode of KBALAP allows the user to retrieve information about a specific event and add or update that event's GT parameters. The program can also create a new event with a GT level for cases where epicenter, time, depth and magnitude GT data are available. Similarly, the batch mode part of the program allows specification of flat files containing GT data for events already in the database.

The RBAP Program

The Regional Body-wave Amplitude Processor (RBAP) is a software tool developed by the Ground-based Nuclear Explosion Monitoring (GNEM) group at Lawrence Livermore National Laboratory (LLNL). RBAP helps to automate the process of making amplitude measurements of regional seismic phases for the purpose of calibrating seismic discriminants at each station. RBAP generates station-centric raw and MDAC corrected Pn, Pg, Sn and Lg amplitudes along with their associated calibration parameters (e.g. phase windows, MDAC values, reference events, etc.) in database tables. It strictly follows the Working Group (WG) 2 standardized processing described in the MDAC White Paper (Walter et al. 2003) and it replaces the original collection of LLNL scripts described by Rodgers (2003). RBAP has a number of advantages over the previous scripts. It is much faster, significantly easier to use, scales more easily to a larger number of events and permits efficient project revision and updating through the database.

RBAP integrates the functions of the modules in the previous LLNL scripts into a single program that is designed to perform the amplitude measurement task efficiently and to require a minimum effort from the users for managing their data and measurements. For well-located events with pre-existing analyst phase picks, the user reviews for quality control and then generates all the amplitudes with just a few mouse clicks. For events needing more attention, the user has complete control over the process (e.g. window control, ability to mark bad data, define regions, define MDAC parameters and define the events to be used in the overall calibration process). RBAP shortens the time needed by the researcher to calibrate each station while simultaneously allowing an increase the number of events that can be efficiently included. RBAP is fully integrated with the LLNL research database. Data is always read directly from the appropriate tables in the research database rather than from a snapshot as was done in the previous system. All RBAP result tables have integrity constraints on the columns with dependencies on data in the LLNL research database. This design makes it very difficult for results produced by RBAP to be stale and also ensures that as the research database expands, RBAP automatically becomes aware of new data that should be processed. RBAP initial users will be LLNL WG 2 members working on IRP's for FY04.
Some RBAP Key Features

• *Based on WG 2 Standardized Algorithm*
-RBAP is built on the WG 2 standardized body-wave amplitude measurement algorithms, documented in the "MDAC White Paper" (Walter et al., 2003). Its results are completely consistent with the last version of the LLNL scripts (Rodgers, 2003) that were vetting in the February 2003 WG 2 exercise between LLNL, LANL and AFTAC.

• *Fast and Efficient Calibration*
-RBAP is self-contained and optimized for station-centric body-wave processing. "Good" events can be handled with just few mouse clicks. The researcher has direct control over key calibration parameters within the tool such as phase amplitude windows and migration, marking bad segments, defining distinct geophysical regions, event types to process, etc. We expect RBAP to provide roughly a factor of 5 increase in calibration speed compared with the original scripts, enabling us to calibrate more stations, with more events per station.

• *Project Management*
- RBAP is designed so that a calibration project can be put down for a day, month or a year, and easily picked up, by the same researcher or a new one. All processing metadata is saved and events are easily tracked as processed, unprocessed or outside the current project definitions. This allows a researcher to efficiently work through a huge data list without repetition and to easily identify and incorporate new events as they become available in the database.

• *Utilizes Database for Up-to-Date results*
- RBAP can draw on the latest calibration parameters being generated by other Working Groups, such as the most recent phase picks, re-locations, magnitudes, instrument response information, or event type ground truth.

• *Batch Processing*
- RBAP is designed to allow simple batch updating of the amplitude results, whether the change is small (e.g. one-event is re-located) or large (instrument response is changed affecting all events).

Database Centric Coordination Framework

As part of our effort to improve our efficiency we have realized the need to allow researchers to easily share their results with one another. For example, as the location group produces GT information, that information should become available for other researchers to use. Similarly, phase arrival picks made by any qualified user should also become immediately available for others to use. This concept extends to sharing of information about data quality. It should not be necessary for multiple researchers to have to repeatedly reject the same bad data. Rather, once data are rejected because of quality reasons they should automatically be excluded from processing by all tools. We are implementing this system behavior using database tables, triggers, stored procedures, and application logic. Although we are at the beginning of this implementation, we have made significant progress over the last year with several kinds of information sharing using the new Database Centric Coordination Framework. These are discussed below.

Significant software engineering and development efforts have been applied successfully to construct an object-oriented database framework that provides database centric coordination between scientific tools, users, and data (Figure 3). A core capability this new framework provides is information exchange and management between different specific calibration technologies and their associated automation tools such as Seismic Location (e.g. KBALAP), seismic identification (e.g. RBAP), and data acquisition / validation (e.g. KBITS). A relational database (ORACLE) provides the current framework for organizing parameters key to the calibration process from both Tier 1 (raw parameters such as waveforms, station metadata, bulletins etc) and Tier 2 (derived measurements such as ground-truth, amplitude measurements, calibration and uncertainty surfaces etc). Efforts are underway to augment the current relational database structure with semantic graph theory structured queries for handling complex queries. Seismic Calibration Technologies (Location, Identification, etc) are connected to parameters stored in the relational database by an extensive object-oriented multi-technology software framework (Figure 3 – middle) that include elements of schema design, Stored procedures, real-time transactional database triggers, constraints, as well as

coupled Java and C++ software libraries to handle the information interchange and validation requirements. This software framework provides the foundation upon which current and future seismic calibration tools may be based.

Sharing of derived event parameters

We have long recognized the inadequacies of the CSS3.0 origin table to serve as a source of information about the "best" parameters for an event. One origin solution may have the best epicenter but poor information on other parameters.
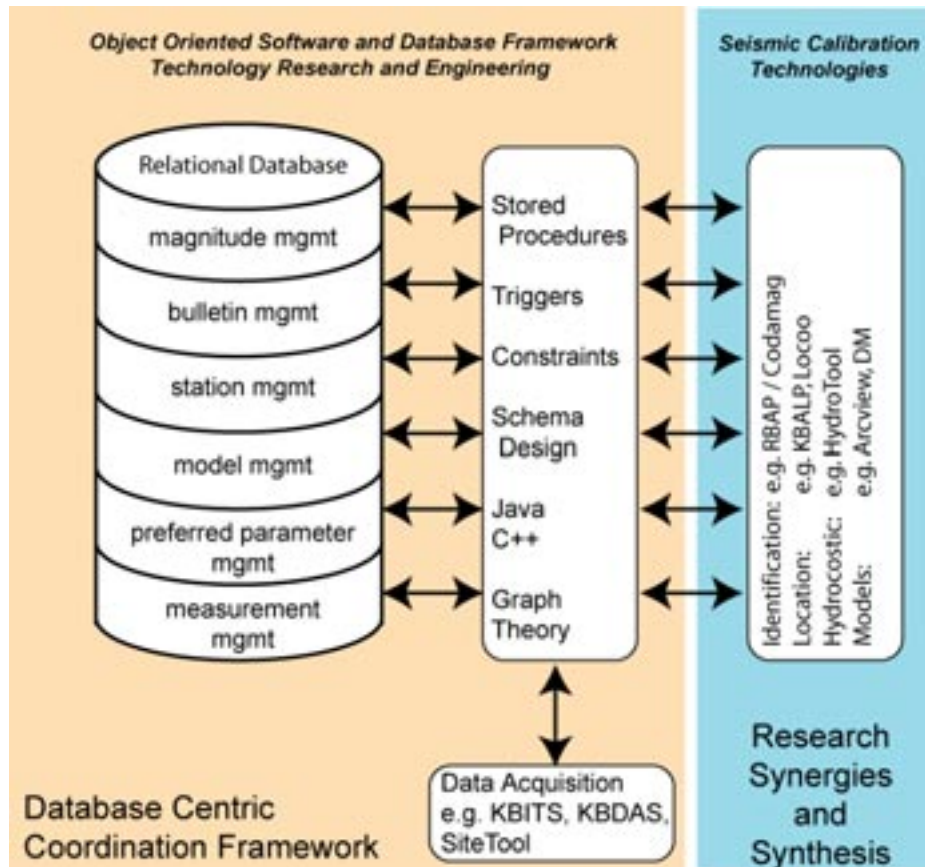

Figure 3. Overview of the Database Centric Coordination Framework that provides the enabling information technology to allow synergy and synthesis of data and calibration technologies for the efficient production of calibration deliverables.

Another may have the correct event type, but be poor in other respects, and so on. We had discussed producing origin table entries with our organization as the author, but that approach has difficulties. Different groups would have responsibility for different fields in the origin. Because their information would not be produced in synchronization we would either have to always be updating the preferred origin or else producing new preferred origins. Also, there would be difficulties in tracking the metadata associated with each field of the preferred origin. Our solution was to create a set of new tables and associated stored procedures and triggers that collectively maintain the "best" information about events. The tables involved are shown in Figure 4. Any tool that needs event parameters references the preferred origin table. It has five parameters of interest:
   1) Event Epicenter
   2) Event Time
   3) Event Depth
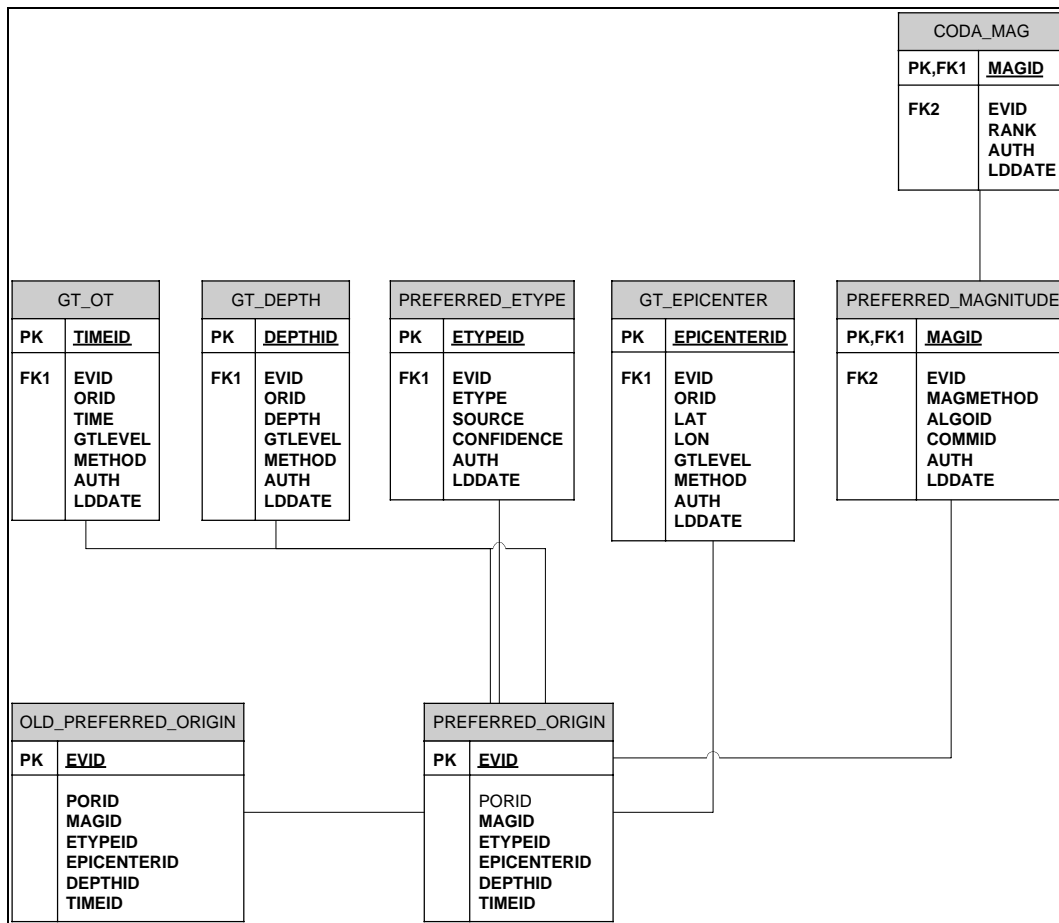   4) Event Moment Magnitude
   5) Event Type.

Figure 4. Tables involved in management of preferred event parameters.

Each of these parameters may be set either from the corresponding columns in the preferred origin from the main origin table or from a table containing the parameter value, confidence level, and other metadata. Rows can be added interactively or using software tools. Inserts, updates or deletes on the main origin table or on any of these tables trigger automatic updating of the preferred origin table. Thus the preferred origin table always has the most current "best" parameter estimates. Whenever a row in the preferred origin table is changed or deleted, a copy of that row is created in the old preferred origin table and the PORID key is updated in the preferred origin table. Thus measurements can always be referenced to the event parameters in existence when the measurements were made.

Currently, both magnitudes and epicenters are being updated automatically as origins enter the system. When the coda magnitude loading tool adds a new coda magnitude, that estimate is automatically ranked against pre-existing coda magnitude estimates, and if it outranks the others (or there are no others) that magnitude becomes the preferred origin magnitude. Whenever a NETMAG entry or an EQ_SOURCE_PARAMS entry occurs, a trigger executes that converts to Mw if necessary, ranks the entry against other Mw estimates for the event, and possibly makes the new estimate the preferred Mw for the event. Whenever a new origin with arrivals enters the system, a set of rules is applied that determine the GT-level (if applicable) of the new origin. If the origin has a GT-level that outranks the current epicenter GT-level, then the epicenter from the new origin solution becomes the preferred epicenter for the event. When the KBALAP tool is complete, it will allow managed changes to preferred origin time, epicenter, depth and etype fields. Currently, the RBAP tool is being modified to read from the PREFERRED_ORIGIN table and to be able to update GT_DEPTH, PREFERRED_ETYPE, and GT_EPICENTER under certain restrictions.

Sharing and Updating of Arrival Information

Although our program has long been using arrivals stored in our CSS3.0 ARRIVAL table, there have been a number of problems that prevented full utilization of those arrivals. These include:

1) Efficiently accumulating all arrivals associated with an event
2) Dealing with the many variations of phase names found in the global bulletins
3) Selecting a "best" pick when more than one author has picked a phase for the same event at some station
4) Associating appropriate site data with each phase pick
5) Where applicable, retrieving filter data associated with the pick
6) Avoiding retrieval of picks previously found to be erroneous.

Again, our solution was to create a set of new tables, stored procedures, functions, and triggers.
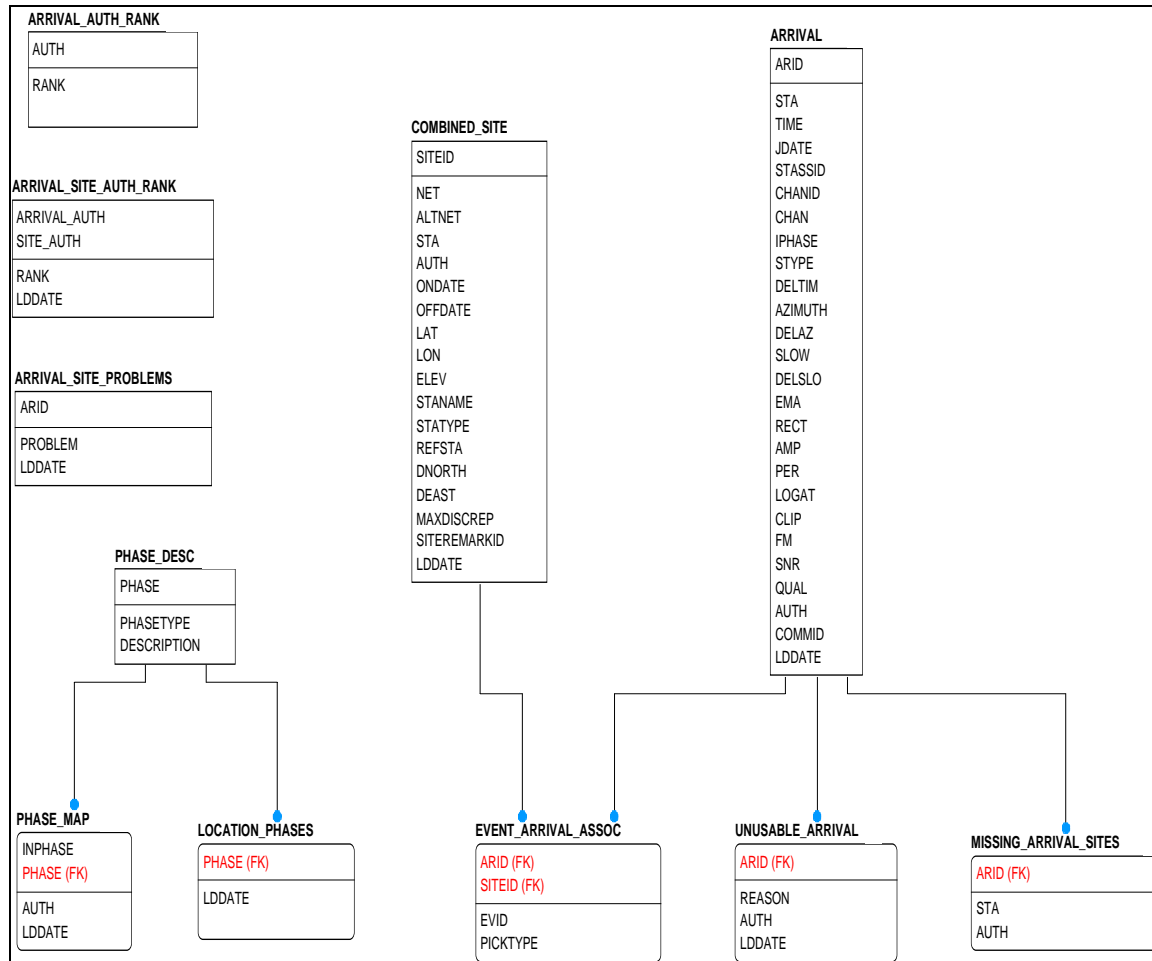
Figure 5. Tables involved in managing arrival information.

The EVENT_ARRIVAL_ASSOC table plays a central role in managing arrivals. In part, it is analogous to the CSS3.0 ASSOC table. But, instead of associating arrivals to origins, it associates arrivals to events. This saves doing a three-table join when selecting all arrivals for an event, an important factor when the arrival table is extremely large. Just as important as being able to easily select all arrivals for an event is the ability to associate the best available site information with each arrival. Arrivals in our system come from a number of different sources, and the associated site information may be inconsistent between authors. This is where the COMBINED_SITE table comes into play. It contains all the site information available to us along with the means to discriminate among sites provided by different authors and for different networks. During loading of arrivals or loading of sites, triggers execute that determine the "best" association between arrivals and sites. This is done using lookups into the

ARRIVAL_SITE_AUTH_RANK table. The association is recorded in the EVENT_ARRIVAL_ASSOC table using the SITEID column.

Retrieval of a set of arrivals and associated sites is accomplished using a stored function returning a REF_CURSOR. This function
- maps phase names to a consistent set using the PHASE_MAP table,
- filters picks for phase suitability using the LOCATION_PHASES table,
- filters picks that have previously been found to be erroneous using the UNUSABLE_ARRIVAL table,
- returns the "best" pick when multiple authors have a pick for the same station-phase,
- and returns the "best" site for each station in the pick set.

The function also allows filtering on the MAXDISCREPANCY column in the COMBINED_SITE table. This, in effect allows rejection of phase picks from stations whose location is not agreed upon to within the user's requested tolerance.

Currently, the RBAP, KBALAP, and TOGEOTOOL programs use this stored function when requesting sets of arrivals. This summer, both KBALAP and RBAP will be given the capability of adding to the UNUSABLE_ARRIVALS table.

Two other tables not shown in Figure 5 are the APPLIED_FILTER and STORED_FILTER tables. The STORED_FILTER describes a particular filter and the APPLIED_FILTER relates an arrival to a stored filter. Currently, these tables are populated by the RTLOADER program when it loads picks made by the GEOTOOL program. This summer RBAP will be given the capability to both read and write these tables. This will allow easy reproduction of the state used to make a certain phase pick, and should be a time saver when reusing picks made by others.

Sharing of QC Information
We have made a start at sharing quality control information among processing tools. So far, the only tool that is writing QC information is RBAP. Users can identify bad segments in traces being processed and write the information into a BAD_SEGMENTS table. This table is visible to all other tools, so it could be used to filter waveforms if appropriate. RBAP also identifies instrument responses that failed, and station-channel-epochs without instrument responses. It is expected that we will write a site metadata tool soon that would, among its other capabilities, use these tables to resolve these problems. In a similar manner, RBAP populates a table of evids for which coda magnitudes need to be computed. At this time, there is still a manual step required in reading this table and using the information to drive the generation of coda magnitudes. When we complete the coda magnitude processing tool, that step should become automatic.

<u>CONCLUSIONS</u>
We present an overview of our software automation efforts and framework to address the problematic issues of very large datasets and varied formats utilized during seismic calibration research and the attributes required to construct next generation data acquisition. These new software and scientific automation initiatives could directly support our current mission including rapid collection of raw and contextual seismic data used in research, provide efficient interfaces for researchers to measure/analyze data, and provide a framework for research dataset integration. The initiatives would improve time critical data assimilation and coupled modeling / simulation capabilities necessary to efficiently complete seismic calibration tasks. The scientific automation engineering and research will need to provide the robust hardware, software, and data infrastructure foundation for synergistic GNEM R&E Program calibration efforts.

<u>REFERENCES:</u>
Ruppert, S., T. Hauk, J. O'Boyle, D. Dodge, M. Moore, Lawrence Livermore National Laboratory's Middle East and North Africa Research Database, Proceedings of the 21st Seismic Research Symposium: Technologies for Monitoring The Comprehensive Nuclear-Test-Ban Treaty, September 1999

O'Boyle, J., S. Ruppert, T. Hauk, D. Dodge, M. Ganzberger, F. Ryall, Middle East, North Africa and Western Eurasia Seismic Research Database, Proceedings of the 25th Seismic Research Symposium: Nuclear Explosion Monitoring – Building the Knowledge Base, September 2003