LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-PROC-414514

Enhancing Seismic Calibration Research Through Software Automation and Scientific Information Management

S. Ruppert, D. A. Dodge, M. D. Ganzberger, D. B. Harris, and T. F. Hauk

July 8, 2009

Monitoring Research Review
Tuscon, AZ, United States
September 20, 2009 through September 23, 2009

ENHANCING SEISMIC CALIBRATION RESEARCH THROUGH SOFTWARE AUTOMATION AND SCIENTIFIC INFORMATION MANAGEMENT

Stanley Ruppert, Douglas A. Dodge, Michael D. Ganzberger, David B. Harris, Teresa F. Hauk

Lawrence Livermore National Laboratory

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

Contract DE-AC52-07NA27344

ABSTRACT

The National Nuclear Security Administration (NNSA) Ground-Based Nuclear Explosion Monitoring Research and Development (GNEMRD) Program at LLNL continues to make significant progress enhancing the process of deriving seismic calibrations and performing scientific integration, analysis, and information management with software automation tools. Our tool efforts address the problematic issues of very large datasets and varied formats encountered during seismic calibration research. New information management and analysis tools have resulted in demonstrated gains in efficiency of producing scientific data products and improved accuracy of derived seismic calibrations.

In contrast to previous years, software development work this past year has emphasized development of automation at the data ingestion level. This change reflects a gradually-changing emphasis in our program from processing a few large data sets that result in a single integrated delivery, to processing many different data sets from a variety of sources. The increase in the number of sources had resulted in a large increase in the amount of metadata relative to the final volume of research products. Software developed this year addresses the problems of:

- Efficient metadata ingestion and conflict resolution
- Automated ingestion of bulletin information
- Automated ingestion of waveform information from global data centers.
- Site Metadata and Response transformation required for certain products

This year, we also made a significant step forward in meeting a long-standing goal of developing and using a waveform correlation framework. Our objective for such a framework is to extract additional calibration data (e.g. mining blasts) and to study the extent to which correlated seismicity can be found in global and regional scale environments.

OBJECTIVE

The NNSA GNEMRD 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 improving the workflow and processing pipeline for seismic calibration products, including the design and use of state-of-the-art interfaces and database centric collaborative infrastructures. These tools must be robust, intuitive, and reduce errors in the research process. This scientific automation engineering and research will provide the robust hardware, software, and data infrastructure foundation for synergistic GNEMRD Program calibration efforts. The current task of constructing many seismic calibration products is labor intensive, complex, expensive and error prone. The volume of data as well as calibration research requirements has increased by several orders of magnitude over the past decade. The increase in quantity of data available for seismic research over the last two years has created new problems in seismic research; data quality issues are hard to track given the vast quantities of data, and this quality information is readily lost if not properly tracked in a manner that supports collaborative research. We have succeeded in automating many of the collection, parsing, reconciliation and extraction tasks individually. Several software automation tools have also been produced and have resulted in demonstrated gains in efficiency of producing derived scientific data products. In order to fully exploit voluminous real-time data sources and support new requirements for time-critical modeling, simulation, and analysis, continued expanded efforts to provide scalable and extensible computational framework will be required.

RESEARCH ACCOMPLISHED

The primary objective of the Scientific Automation Software Framework (SASF) efforts is to facilitate the development of information products for the GNEMRD 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 derived information products for delivery into the NNSA KB.

These next generation information management and scientific automation tools are used together within specific seismic calibration processes to support production of tuning parameters for the United States Atomic Energy Detection System (USAEDS) run by the Air Force. The automation tools create synergy and synthesis between complex modeling processes and very large data sets by leveraging a scalable and extensible database centric framework. The requirements of handling large datasets in diverse formats, and facilitating interaction and data exchange between tools supporting different calibration technologies, has led to an extensive scientific automation software engineering effort to develop an object oriented database-centric framework using proven research-driven workflows and excellent graphics technologies as a unifying foundation.

The current framework supports integration, synthesis, and validation of the various information types and formats required by each of the seismic calibration technologies. 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 geospatially registered correction surfaces and uncertainty surfaces. In past years, our automation efforts have been largely focused on research support tools, RBAP (Regional Body-wave Amplitude Processor) and KBALAP (Knowledge Base Automated Location Assessment and Prioritization) (Ruppert, et al. 2005, Ruppert, et al. 2007, Hauk, et al. 2008). However, this year, we have shifted development effort to address some fundamental changes in the nature and use of the data that is integrated into our system.

Much of our data ingestion infrastructure was written in the late 1990s when seismological data centers were fewer in number and possessed minimal capabilities. At that time, there was little infrastructure to support automated retrieval of data, so our codes all required that the data to be loaded must exist on our intranet. In addition, at the start of the last decade our efforts were devoted solely to the support of the base program objectives and we had only a single analysis tool (SAC) (Ruppert, et al. 1999). Today, we ingest data from many sources for use in creating a variety of research products over and above the calibration products of the base program. The change is shown schematically in Figure 1 below.

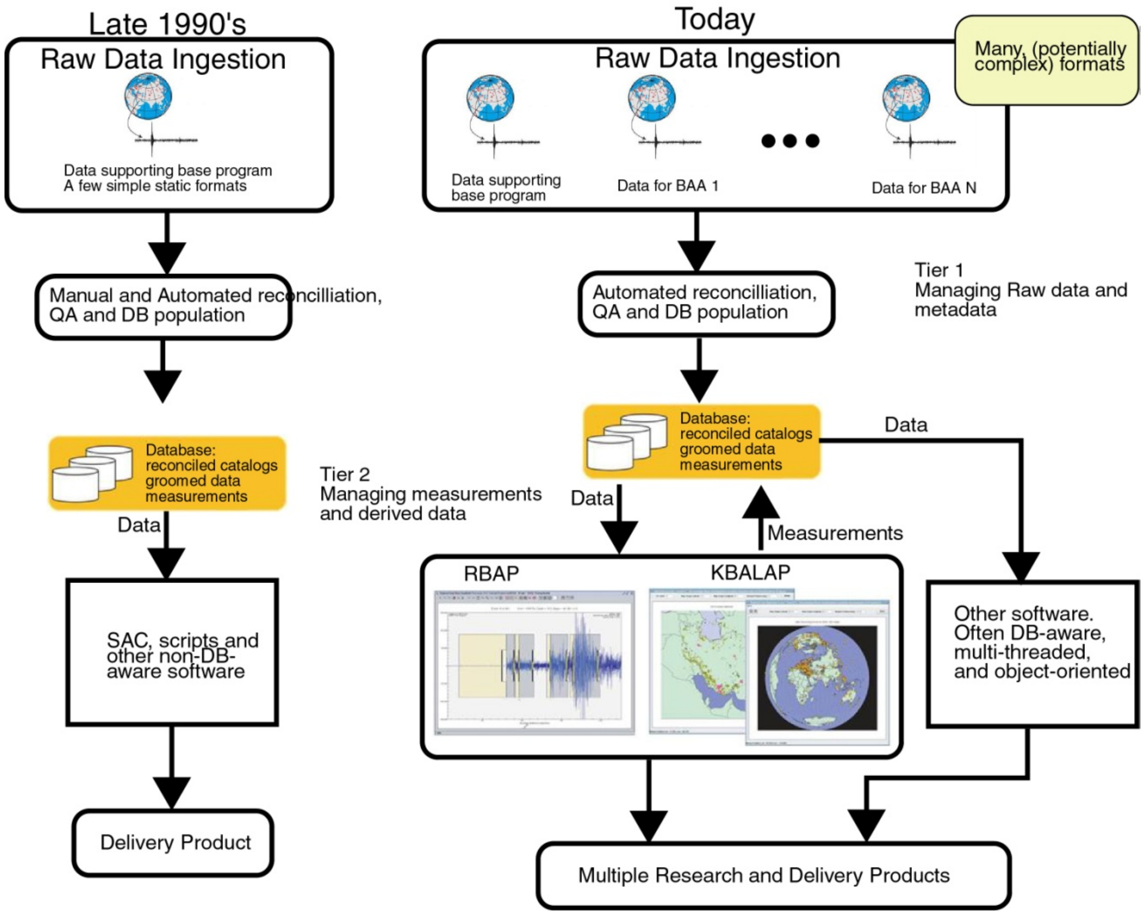

Figure 1 shows a schematic comparison of differences in dataflow between the late 1990s and today. Ten years ago, the bulk of our data was from a few global sources and a common processing was applied to the data to produce a single integrated delivery product. Today, in addition to global data sources, data from scores of local and regional networks flow into our system. Many different types of processing may be applied to the data to produce a number of research products.

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:

<u>Tier 1: Contextual and Raw Data</u>
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

Although we have taken advantage of many of the new sources of data as they became available, for the most part our data ingestion has retained a multistep process. The sequence has generally been: 1) Obtain data via FTP, EMAIL, TAPE or other physical media, 2) Stage Data and 3) Use variety of automated loading tools to add data to database.

At the beginning of our automation efforts in early 2000, we automated each of these individual steps (KBITS) (Ruppert et al., 1999, Ruppert et al. 2004) in order to quickly and efficiently process large volumes of common/simple formatted data. This allowed us to go from 500K to 200M waveforms (Table 1) (Ruppert et al. 1999, Hauk et al. 2008). Examples include global bulletins and data from a few datacenters in known usually invariant formats. Subsequent efforts focused on automation of measurement and metadata production of Tier 2 topics. Tier 2 efforts of our software development effort included production and refinement of analysis tools such as RBAP, KBALAP, GT-MERGE [Ruppert et al, 2005, Ruppert et al., 2007, Hauk et al, 2008].

Although software automation (KBITS) made each individual step of loading Tier 1 data more efficient, further gains in efficiency and capability can be achieved through better integration and consolidation of individual steps. This approach also applies for acquisition and loading of both bulletins and waveforms. Channel data have been mostly obtained from SEED files and from CSS flat files. However, over time there has been a slow shift in the nature of the data that our researchers require, and this is putting increasing strain on our semi-automated data loading efforts. In the late 1990s, we concentrated most of our efforts on a relatively small number of stations for which we could get data in large blocks. For this scenario, a small amount of manual effort could result in a large number of rows in the WFDISC, ORIGIN, and ARRIVAL tables. Today, with the increased importance of BAA (Broad Area Announcement) research projects in our program, we are processing many more relatively small data sets. So, the proportion of time spent loading metadata per waveform row loaded has increased substantially.

1999 Status	**2009 Status**
• 0.1 TB storage	• 200 TB storage
• 10K Events	• 3M events
• 500K waveforms	• 200M waveforms
• 100K measurements	• 750M measurements
• Single Program	• Multiple Programs
• Single Data Domain	• Multiple Data Domains
• Single Tool (SAC)	• Multiple Tools (RBAP, KBALAP, GTMERGE, KBITS, Delivery Schema)
	• Uncertainties/Validation/Metadata
	• GUI/GIS/Mapping/Visualization
	• Support calibration efforts such as GTMerge and EventID deliveries
	• Support multiple BAA projects

Table 1 shows the change in complexity of our data processing environment in terms of data volumes and display and output requirements.

It is also the case that changing metadata can be much more difficult today than it was several years ago. This is mostly because a large number of rows in many tables can be affected by a change to metadata. It is more difficult to keep track of all the potentially affected data, and it can be time consuming just to update the affected rows.

Meanwhile, seismological data centers have grown in number and matured in sophistication to the point where it appears feasible to get a significant fraction of the data into our system automatically. Based on all these factors, we decided that the time was right to shift some focus back to our data ingestion effort and try to introduce some automations where possible. To that end, we have constructed in the last year, three pilot applications for automating data ingestion. These are: ***UsgsEnsClient, Station Metadata Tool, and Waveform Retriever***

UsgsEnsClient

This tool takes advantage of the USGS ENS Earthquake Notification Service (USGS ENS Documentation, 2009) to get a first-look origin into our system within an hour or so of the event occurrence. Although we continue to load global bulletins as before, by having the event in our system early on, we are able to analyze events of interest much more rapidly than would be the case if the origin information had to be retrieved manually and run through our old bulletin loading system. We do not currently retrieve events for the entire Earth through this mechanism. Instead we have defined several regions of interest for which we load all ENS solutions. Nevertheless, since its inception, this system has reliably loaded over 63,000 events.

Station Metadata Tool

The Station Metadata Tool software is a Java application that builds on our previous work in merging SITE data from multiple sources into a single CSS (Center for Seismic Studies) SITE table. The STATION_METADATA schema and supporting software (Hauk et al, 2008) has proven useful in tracking the provenance of SITE data and making informed choices in the face of conflicting information. However, it is only a partial solution to the problem of maintaining complete up-to-date metadata for our system.

That software was predicated on the idea of periodic (relatively infrequent) manual introduction of new data. When this occurred, the software would analyze the data for inconsistencies, and after they were resolved, the entire SITE table would be rebuilt. As it happens, even the major data sources (NEIC (National Earthquake Information Center) and ISC (International Seismological Center) station books) are revised several times per year. In addition, virtually every week, there is some new set of stations from an unconventional source that is required to be loaded into our database. The SITE rebuilding mechanism is just too cumbersome for this workflow.

Also, getting the SITE information into the database is only part of the problem. Once the SITE data are in place, we are still faced with a largely manual process of building/updating channel and response information. Of course since there are potentially many channels per SITE this process can be quite onerous.

Another issue that we face in loading SITE data is the rather frequent occurrence of conflicting station codes. We are frequently asked to load data obtained from a temporary deployment or from a cooperative agreement with a regional network operator. Often in these cases, some station codes are already in use by the NEIC or else a local code has been adopted for a NEIC station. It is possible to research each station code before attempting to load the station, but done manually, this is a somewhat tedious task. Fortunately, it is a process that can be automated.

There is now a wealth of information about seismic stations and instrumentation on the Web. Probably the most advanced access system for seismic metadata is the DHI (Data Handling Interface) framework currently implemented by the IRIS DMC (Incorporated Research Institutes for Seismology Data Management Center), the University of California at Berkeley Data Center and others. This framework allows direct programmatic access to metadata. With this system it is easy for a client to request: site data, channel data, and instrument response data.

A client can also easily enumerate all the networks and stations known to the server. The principal limitation of this service (at least at the IRIS DMC) is that metadata is only available for stations for which they have archived waveform data.

In addition to IRIS, the NEIC, ISC, and ORFEUS (Observatories and Research Facilities for European Seismology) data centers have station data available via FTP, and static Web pages. For these sources, it is straightforward (although somewhat fragile) to write methods based on Web scraping that provide a programmatic query interface. We have written a Java Interface for finding station metadata and have produced implementations for the above-mentioned data centers. This is one of the core pieces of the new Metadata Tool.

The Metadata Tool is an interactive station metadata search, entry, and editing tool. It introduces a considerable amount of automation into this process to ease the burden on the user. First, for several major sources of station metadata (NEIC, ISC, IRIS, ORFEUS) the tool automatically retrieves SITE data for new or modified stations with a single button click. For a station that a user is considering to enter manually, the tool will identify all known (within its universe of data centers) information about the station. For retrieved codes that are new to our system, and for which there are no nearby stations (either in our system or in one of the external databases), the software automatically populates both the STATION_METADATA schema entries and the LLNL.SITE entry. In the case of conflicts, the software presents all available information to the user allowing the user to resolve the conflict. Steps taken to resolve conflicts are automatically recorded in the STATION_METADATA schema along with user comments. After conflict resolution is complete, the LLNL SITE table is updated with the resulting new information.

For new entries with conflicting codes, the conflict is automatically recognized and the user is prompted to remap the local code to the NEIC code or to a new non-conflicting code as required. In these cases, the software also establishes a context remap that our other data loading software will use when loading waveforms or arrivals for the affected stations.

The Metadata Tool has three options for loading channel and sensor data. For the case where IRIS has a record of the metadata, the user can just select the desired channels from a list and click a button to request the data from IRIS. This causes all necessary SITECHAN, SENSOR, and INSTRUMENT rows to be built and the response file to be installed in the correct location in our file system. For cases where the user has a set of response files, and knows what channels they are to be applied to, the software will build the required tables based on a minimal amount of user input, and will copy the user-supplied response files to the correct place in our file system. The last supported case is a generic option. With this option, the user simply identifies the bands for which instrument responses are required. The software then builds generic response information from stored templates.

Instrument Response Conversion and Database Integration

We collect and maintain metadata for many seismic stations around the globe. Currently, we have over 14,000 distinct station/channel combinations for which we have instrument responses. The vast majority of those responses are from IRIS in SEED format (SEED Reference Manual, 2009). The IRIS RESP format is used for all internal processing and analysis. For certain products we also maintain and produce a PAZFIR format from the original IRIS RESP formatted files. This section describes the conversion framework we developed for accomplishing this task and some of the challenges it addresses.

Because the IRIS RESP format allows for considerable variation in the blockettes that can be present in a response file, and usage of the blockettes has evolved with time, it hasn't always been possible to generate an equivalent PAZFIR for every possible RESP response. Thus, an integral part of our processing framework has been including a mechanism to test all converted responses. The response processing code is written primarily in Matlab utilizing some of the parsing logic previously developed by George Randall and Richard Stead at Los Alamos National Laboratory. The code iterates over a directory of RESP files and for each file it uses the algorithms to convert the response to PAZFIR. Next it calls the Java JEvalResp code [Instrumental Software Technologies, Inc., 2009)] to produce a Frequency-Amplitude-Phase (FAP) file from the original RESP file. The code then executes the Matlab version of the NDC unscaled_response code (Chris Young, personal communication) to produce a second FAP file constructed on the same set of frequencies as the first. Finally the code reads the two FAP files, converts them to complex vectors, and computes the residual. Depending on the size of the residual, the converted response is either accepted or rejected.

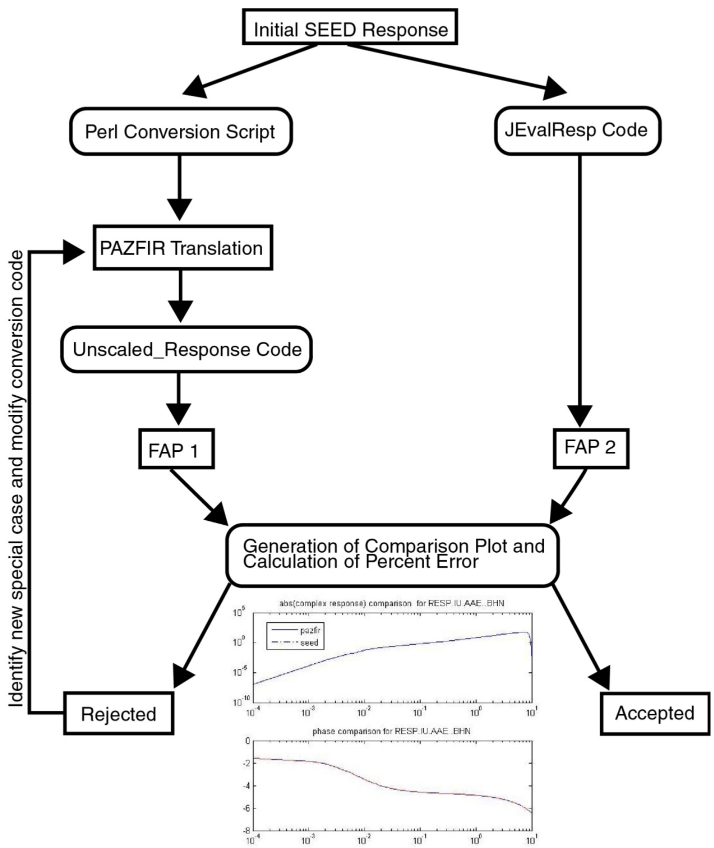

Figure 2. This is a schematic diagram of the response conversion framework that we have been using to convert IRIS RESP format responses into the PAZFIR format.

Once all the responses have been converted to a target format the final step is to integrate them into the target database. If the source response is an IRIS SEED file then it may in fact, hold one or more individual response epochs. If that is the case then special consideration is applied to break those out into individual target responses. As we integrate response epochs in a target database, we must insure that they agree with existing station and channel epochs as well as insure that there are no waveforms for which we miss instrument response coverage. Any discrepancy needs to be addressed to insure we have accurate and consistent metadata across the various target database tables.

Waveform Retriever

The Waveform Retriever is the third piece of automation software that we have developed this year. It is a DHI client whose purpose is to retrieve all available event waveform segments for a defined set of stations and channels.

Retrieved data are merged as required into our WFDISC table. The tool also verifies that all required metadata are in place prior to writing any new WFDISC rows, and will make SITECHAN, SENSOR, and INSTRUMENT rows as required using methods developed for the Metadata Tool. The Waveform Retriever can be used with the IRIS BUD server or with the archive server, so it is suitable for both special event analysis or for routine data loading. Our expectation is that we will develop a framework around this tool that will replace our current semi-automated system with a fully automated system.

Prototype Subspace Detection System

For many years our program has had interest in developing a correlation-based detection system that could operate on our many Terabytes (and growing) collection of continuous and segmented waveforms. In addition to facilitating our understanding of the degree to which correlated seismicity exists on a global scale, the resulting database of categorized event segments could prove quite valuable in improving our travel time and seismic discriminant calibrations.

Primarily because of resource limitations, we have never implemented such a system. An additional factor in delaying our implementation of such a system has been the prohibitive amount of labor that would be required in building and maintaining the set of templates that would be required to process our entire holdings. This year, Dave Harris received funding for a BAA project to investigate the possibility of developing an adaptive correlation detector framework. The goal was to develop a framework capable of creating its own templates with minimal supervision. The system that resulted from this effort was reasonably successful, and we are intending to adapt it to our requirements.

The framework developed for the BAA project can operate on an arbitrary number of channels from one or more stations. The system starts with a single STA/LTA or array power detector. Each detection by that detector results in the creation of a rank-one sub-space detector. For all subsequent segments, the sub-space detectors run in parallel with the STA/LTA detector. This policy of automatically spawning subspace detectors from unique STA/LTA triggers has worked well on the test sets processed so far, in part because the system is very fast. For example, approximately 89,000 seconds of 100 sps, single channel Mt. St. Helens data were processed in 8.5 seconds. The great speed is due in part to the innovative decimation algorithm used [Harris and Paik, 2006].

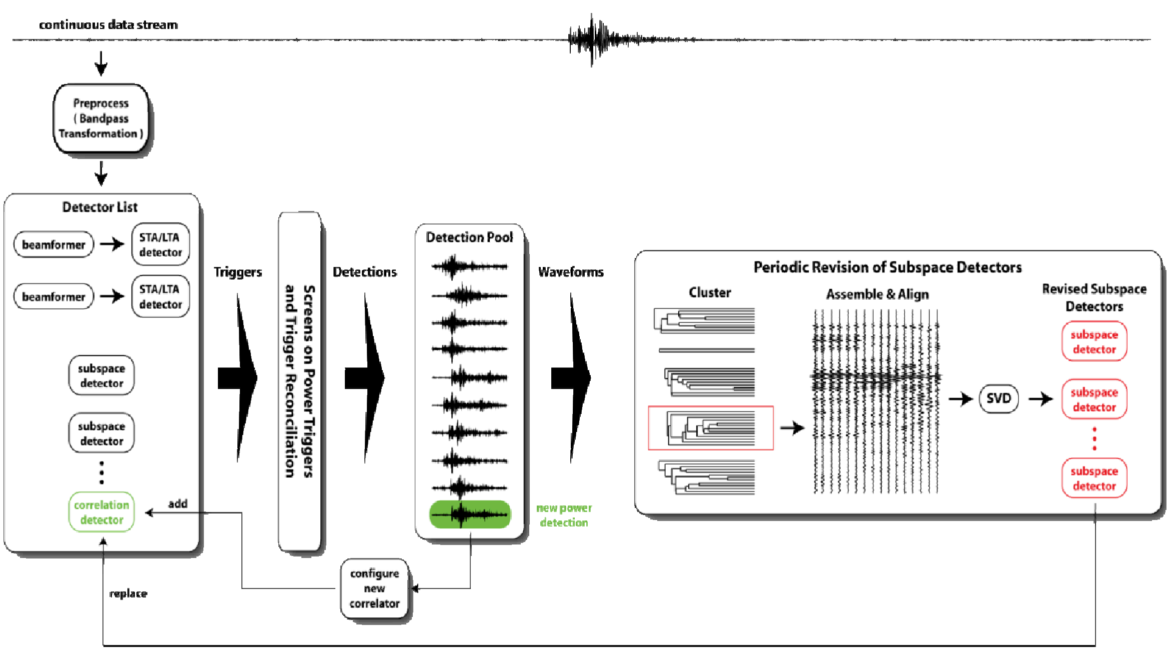

Figure 3 (from **Figure 1** of Harris, 2009) A framework for signal detection consisting of both conventional power detectors and correlation detectors generated automatically from a continuously updated pool of waveform templates. The conventional detectors are permanent members of the list and serve to detect signals for which previous observations do not exist. Periodically the system halts processing of the data stream and replaces the correlation detectors with subspace detectors built from clusters of waveforms in the detection pool. This function tends to check the growth of unlimited numbers of detectors. This research prototype operates without operator intervention, which practical systems would require.

Periodically, a supervisor component halts the detection process in order to coalesce the current set of subspace detectors. The reason for this step is that many signal segments may have waveforms that are similar to the templates for multiple rank-1 subspace detectors. In these cases, multiple detectors will declare detections. This complicates any post-detection association process. We refine the detector collection by extracting the detections for all automatically-spawned detectors since the last refinement. All possible pair-wise correlations are computed for the set of signals followed by a complete-link cluster analysis. For each of the resulting clusters, the signals are aligned using their cross-correlation-derived shifts. For each set of aligned signals, a singular value decomposition (SVD) is calculated and a new subspace basis from rank 1 – 3 is computed. Each basis is used to create a new subspace detector. At that point, all the input detections are assigned to the new detectors and the detectors are added to the framework. Any input detectors left with no detections are then retired after which, the detection process is resumed. This process is outlined in Figure 3 above from (Harris, 2009).

The framework built for the BAA research includes a database-hosted persistence layer that stores triggers, detections, and detector information. The detector information includes not only the parameters under which the detector operates, but also the subspace templates stored as BLOBs. This framework is a good start toward what will be required for deployment in a system to meet our program’s needs. Additions to the framework will include

- Index tables for both continuous and segmented waveforms. These tables will track both availability of data from scanning by detectors, and records of what has already been scanned (Required for restartability and work sharing among threads).
- A mechanism to correlate detections with events identified externally. This will support arrival correlation, identification of mine blast clusters, and general studies of correlated seismicity.
- A monitoring and dispatching system that is responsible for controlling detector threads, viewing the status of active detectors, and accessing cluster statistics.

CONCLUSIONS AND RECOMMENDATIONS

We present an overview of our software automation efforts and framework to address the problematic issues of consistent handling of the increasing volume of data, collaborative research efforts and researcher efficiency, and overall reduction of potential errors in the research process. By combining research driven interfaces and workflows with graphics technologies and a database centric information management system coupled with scalable and extensible cluster based computing, we have begun to leverage a high performance computational framework to provide increased calibration capability. These new software and scientific automation initiatives will directly support our current mission including rapid collection of raw and contextual seismic data used in research, provide efficient interfaces for researchers to measure and analyze data, and provide a framework for research dataset integration. The initiatives will improve time-critical data assimilation and coupled modeling and simulation capabilities necessary to efficiently complete seismic calibration tasks. This GNEMRD Program's scientific automation, engineering and research, will provide the robust hardware, software, and data infrastructure foundation for synergistic calibration efforts.

ACKNOWLEDGEMENTS

We acknowledge the assistance of the LLNL computer support unit in implementing and managing our computational infrastructure. We thank Jennifer Aquilino, Laura Long, and John Hernandez for their assistance in configuration and installation of our Linux cluster and HPC, Storage, and desktop workstations.

REFERENCES

Elliott. A, , D. Dodge, M. Ganzberger, T. Hauk, E. Matzel, and S. Ruppert (2006), Enhancing Seismic Calibration Research Through Software Automation and Scientific Information Management, *Proceedings of the 28th Seismic Research Review: Ground-Based Nuclear Explosion Monitoring Technologies*, 28, 967-975

Harris, David B. (2008), Covariance Modifications to Subspace Bases, Lawrence Livermore National Laboratory, internal report, xxxxxxxx, xxpp.

Harris, D. B. (2009), Adaptive Waveform Correlation Detectors for Arrays: Algorithms for Autonomous Calibration in current Monitoring Research Review proceedings, pp. TBD.

Harris, D. B. and T. Paik (2006), Subspace Detectors: Efficient Implementation, Lawrence Livermore National Laboratory, internal report UCRL-TR-223177, 38 pp.

Hauk, T.F., D.A. Dodge, M.D. Ganzberger, E.M. Matzel, S.D. Ruppert (2008). Enhancing Seismic Calibration Research through Software Automation and Scientific Information Management, in *Proceedings 30th Monitoring Research Review: Ground-based Nuclear Explosion Monitoring Technologies*, pp. 978-987.

ISTI—Instrumental Software Technologies, Inc. Home Page, 2 July 2009, <http://www.isti.com>.

Ruppert S., D. Dodge, M. Ganzberger, T. Hauk, E. Matzel (2007), Enhancing Seismic Calibration Research Through Software Automation and Scientific Information Management, in *Proceedings of the 29th Monitoring Research Review: Ground-Based Nuclear Explosion Monitoring Technologies*, LA-UR-07-5613, Vol 1, pp. 965-974

Ruppert, S., T. Hauk, J. O'Boyle, D. Dodge, and M. Moore (1999), 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*, 21, 243

SEED Reference Manual
Standard for the Exchange of Earthquake Data
SEED Format Version 2.4
January, 2009, 2 July 2009, < http://www.iris.edu/manuals/SEEDManual_V2.4.pdf>.

US Geological Survey Earthquake Notification Services
USGS ENS Documentation
Page Last Modified: January 06, 2009 18:04:11 UTC, 2 July 2009
<https://sslearthquake.usgs.gov/ens/?page=help>.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.