SAND2000-1952C

REDUCING INFORMATION OVERLOAD IN LARGE SEISMIC DATA SETS

Jeff Hampton, Chris Young, John Merchant, Dorthe Carr and Julio Aguilar-Chang[1]

Sandia National Laboratories and [1]Los Alamos National Laboratory

Sponsored by U.S. Department of Energy
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation
National Nuclear Security Administration

Contract No. DE-AC04-94AL85000

ABSTRACT

Event catalogs for seismic data can become very large. Furthermore, as researchers collect multiple catalogs and reconcile them into a single catalog that is stored in a relational database, the reconciled set becomes even larger. The sheer number of these events makes searching for relevant events to compare with events of interest problematic. Information overload in this form can lead to the data sets being under-utilized and/or used incorrectly or inconsistently. Thus, efforts have been initiated to research techniques and strategies for helping researchers to make better use of large data sets. In this paper, we present our efforts to do so in two ways: 1) the Event Search Engine, which is a waveform correlation tool and 2) some content analysis tools, which are a combination of custom-built and commercial off-the-shelf tools for accessing, managing, and querying seismic data stored in a relational database.

The current Event Search Engine is based on a hierarchical clustering tool known as the dendrogram tool, which is written as a MatSeis graphical user interface. The dendrogram tool allows the user to build dendrogram diagrams for a set of waveforms by controlling phase windowing, down-sampling, filtering, enveloping, and the clustering method (e.g. single linkage, complete linkage, flexible method). It also allows the clustering to be based on two or more stations simultaneously, which is important to bridge gaps in the sparsely recorded event sets anticipated in such a large reconciled event set. Current efforts are focusing on tools to help the researcher winnow the clusters defined using the dendrogram tool down to the minimum optimal identification set. This will become critical as the number of reference events in the reconciled event set continually grows. The dendrogram tool is part of the MatSeis analysis package, which is available on the Nuclear Explosion Monitoring Research & Engineering Program Web Site (http://www.ctbt.rnd.doe.gov/ctbt/data/matseis/matseis.html).

As part of the research into how to winnow the reference events in these large reconciled event sets, additional database query approaches have been developed to provide windows into these datasets. These custom built content analysis tools help identify dataset characteristics that can potentially aid in providing a basis for comparing similar reference events in these large reconciled event sets. Once these characteristics can be identified, algorithms can be developed to create and add to the reduced set of events used by the Event Search Engine. These content analysis tools have already been useful in providing information on station coverage of the referenced events and basic statistical information on events in the research datasets. The tools can also provide researchers with a quick way to find interesting and useful events within the research datasets. The tools could also be used as a means to review reference event datasets as part of a dataset delivery verification process. There has also been an effort to explore the usefulness of commercially available web-based software to help with this problem. The advantages of using off-the-shelf software applications, such as Oracle's WebDB, to manipulate, customize and manage research data are being investigated. These types of applications are being examined to provide access to large integrated data sets for regional seismic research in Asia. All of these software tools would provide the researcher with unprecedented power without having to learn the intricacies and complexities of relational database systems.

Key Words: information overload, Event Search Engine, winnowing, dendrogram, tools, database, WebDB.

OBJECTIVE

The main objective of our research is to provide tools and techniques that can help seismic event monitoring researchers make better use of large event catalogues. One part of our effort focuses on developing a method to identify similar waveforms in a large archived set. Another aspect of our research involves improved methods for accessing event catalog data stored in a relational database system. We feel that proper use of relational database systems will greatly improve the researchers' ability to utilize and understand the data in their catalogs.

RESEARCH ACCOMPLISHED

Below we discuss three technologies that have been developed or are being researched in order to fulfill the objectives described above. These are: 1) Event Search Engine, 2) Content Analysis Tools, and 3) WebDB.

1) *Event Search Engine*:

We are developing a set of tools known as the Event Search Engine (ESE) to provide a way to search large seismic data sets in an optimized and faster way based on a waveform correlation. The use of waveform correlation to identify similar events is well-established (e.g. Israelsson, 1990; Harris, 1991; Aster and Scott, 1993; Riviere-Barbier and Grant, 1993), but there are problems with applying it to large data sets. If a researcher has a catalog of several thousand events or more, the time it would take to compare all the catalog waveforms with a waveform of interest could be prohibitive, at least for common use. The ESE is being developed to deal with this problem.

One component of the ESE is a hierarchical clustering tool known as the dendrogram tool, which is written as a MatSeis graphical user interface (GUI). The dendrogram tool allows the user to build dendrogram diagrams for a set of waveforms, which would typically be an archived set plus one from a current event of interest. These diagrams can be used to establish any groupings of similar waveforms within the set. Dendrograms are built as an iterative process in which the most similar set of waveforms is found first and formed into a group, and the next most similar pair is found, with the group considered in the same way as individual waveforms. The trick to this method is in the way the correlations are calculated between the individual waveforms and the group. There are many methods to do this (e.g. single linkage, complete linkage, flexible method – see Krzanowski, 1988), and our tool allows the user to choose which method they want to use. Prior to calculating the dendrogram, the user can apply various signal-processing parameters to improve the correlation including phase windowing, down-sampling, filtering, and enveloping. We also allow the clustering to be based on two or more stations simultaneously, which is important to bridge gaps in sparsely recorded regional events. Based on the settings the user selects, dendrograms can be quickly and easily generated for a variety of clustering techniques in combination with different waveform filtering and down-sampling choices. This allows the researcher to find the optimal combination for his research goals.

The dendrogram picture that is produced as a result of the waveform comparison shows the researcher in a concise visual way the clusters of similar waveforms. (See Figure 1) The waveform of interest that was selected is highlighted so the researcher can see at a glance how it fits in with the other waveforms. For example, if the waveform of interest joins to a group at a low correlation level, then there is a clear indication that the waveform of interest is not like any of the other waveforms compared.

DISCLAIMER

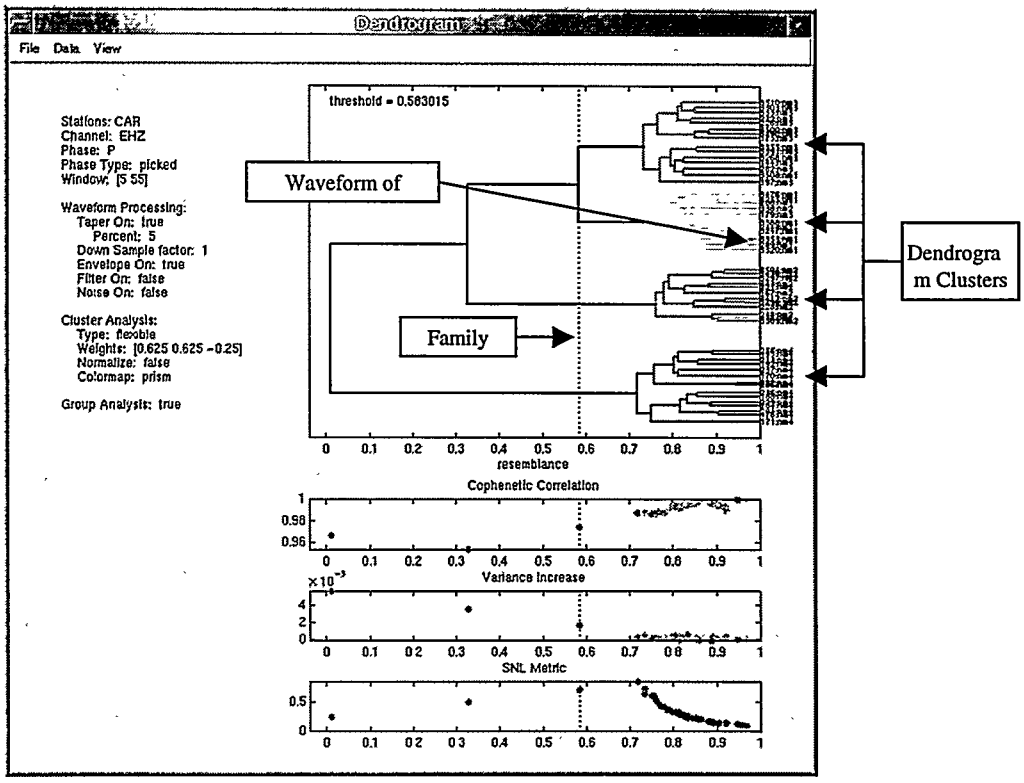


Figure 1. Example Dendrogram Cluster.

Perhaps the least understood aspect of dendrogram analysis is where to draw the vertical line that establishes the resulting clusters, which we call families. Our default location for this line is where the largest drop between successive cophenetic correlations occur (Ludwig and Reynolds, 1988). Cophenetic correlation is a measure of the correlation between the actual waveform correlations and those predicted by the dendrogram. A cophenetic correlation value can be calculated after each join in the iterative dendrogram formation process, comparing actual and theoretical values for only the waveforms that have been clustered to that point. Note that cophenetic correlations must be exactly 1.0 until a waveform is joined to a group, because the actual and theoretical correlations of a waveform to another waveform are the same. As a method of automatically establishing the family line, we have found that cophenetic correlation works a relatively small amount of the time. However, we continue to use it because we have found no better method and because the plot of cophenetic correlation sometimes proves useful in understanding the dendrogram. In fact, the automatically chosen location of the family line is not particularly important as the user can move the line to any position and see how this affects the families. To make this clearer, whenever the line is moved, we assign a different color to each family to the right of the line and redraw the dendrogram.

This coloring is also used for plots of the waveforms and of the locations of the events. The waveforms can be sent to the MatSeis utility Freeplot, where they will appear in the same order as in the dendrogram with each waveform correctly shifted for the best correlation with the waveform immediately above it. (See Figure 2) This allows the user to quickly evaluate whether the families in the dendrogram are grouping waveforms with useful characteristics. If the data is sent to the map, the result is a plot of the locations of each event with the symbol assigned the same color as from the dendrogram. (See Figure 3) This is a very useful option which, when combined with the waveform plot, can be used to evaluate the locations of a set of similar events. To facilitate this, the user can draw a box around any set of origins in the map and highlight this same set back in the dendrogram to see whether or not they have been grouped.

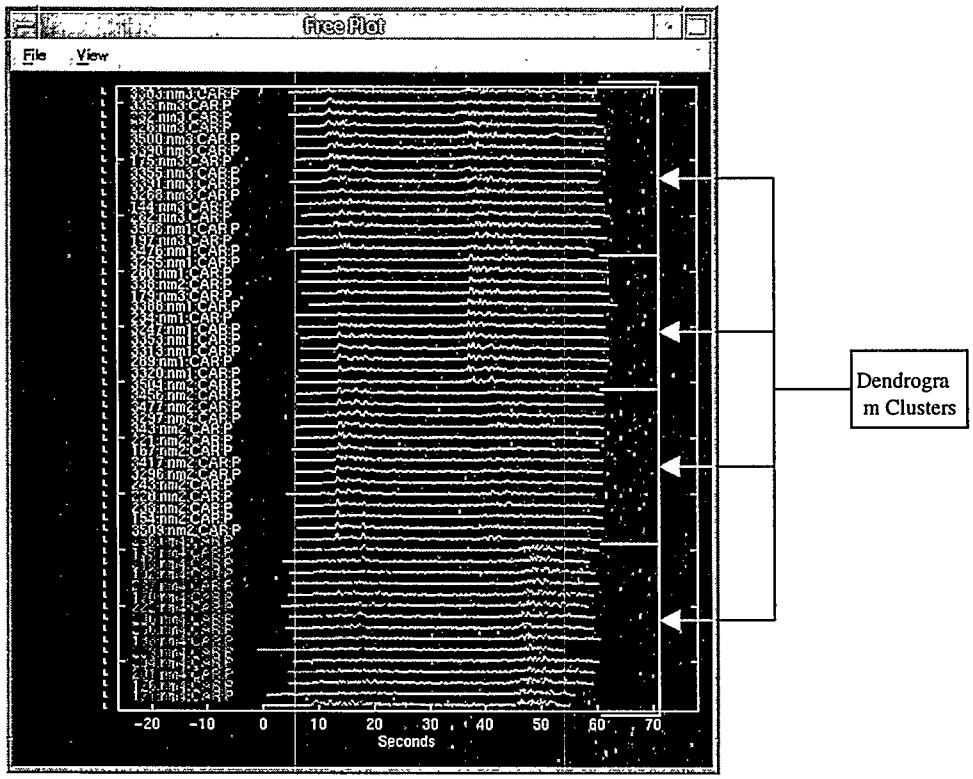


Figure 2. Freeplot Sample Screen.

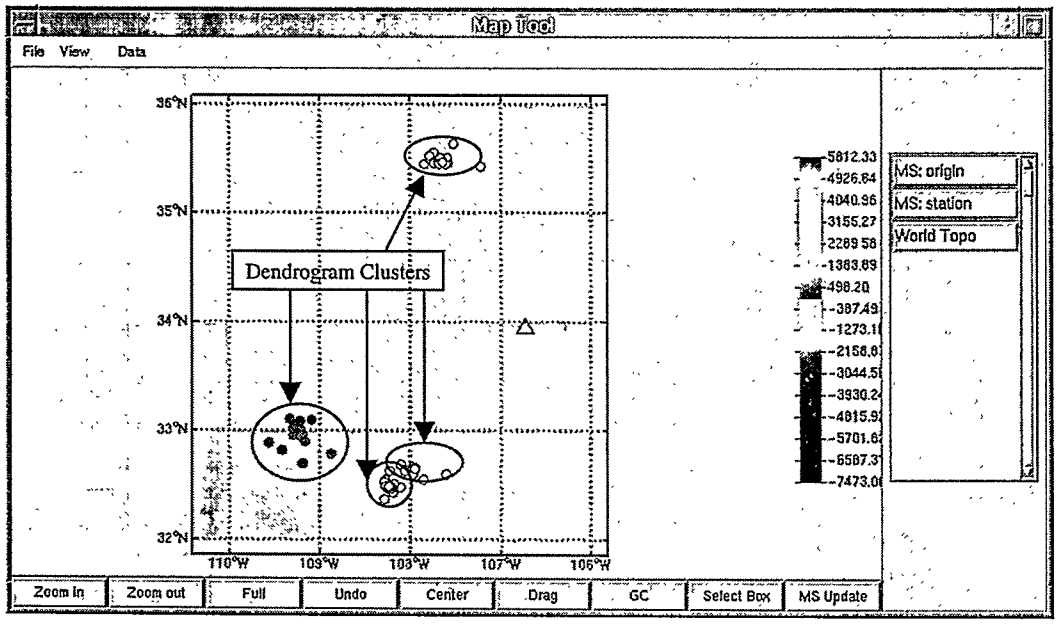


Figure 3. Map Tool Sample Screen.

2) *Content Analysis Tools*:

One of the problems faced by us in order to begin using the dendrogram tool with large data sets was how to find the stations with good waveform coverage, which would lead to meaningful dendrograms. Requesting a set of waveforms for a large list of origins is a time-consuming process even if they are managed with a relational database,

so proceeding with one station after another can be tedious. A better solution is to provide some means to see which of the stations is best represented in the data set, and to work on these first. This was one of the springboards that started us thinking about ways of visualizing summary information about the data sets that we were using. If we had a dynamic way of summarizing the content of our data sets, we could considerably improve the process of creating a dendrogram. Therefore, several web-based tools were developed to produce detailed listings and several choices of charts that would help us to understand our data sets better.

The web-based tools we developed allow the user to query the database data sets based on several criteria such as station, channel, latitude range, longitude range, and date range. (See Figure 4) These tools also distinguish between querying against origins with waveforms and origins with arrivals. These content analysis tools allow the researcher to quickly visualize their data sets given a set of user-supplied values. All the while, the researcher hasn't had to deal with the database directly or learn SQL, a database query language, to formulate their queries.

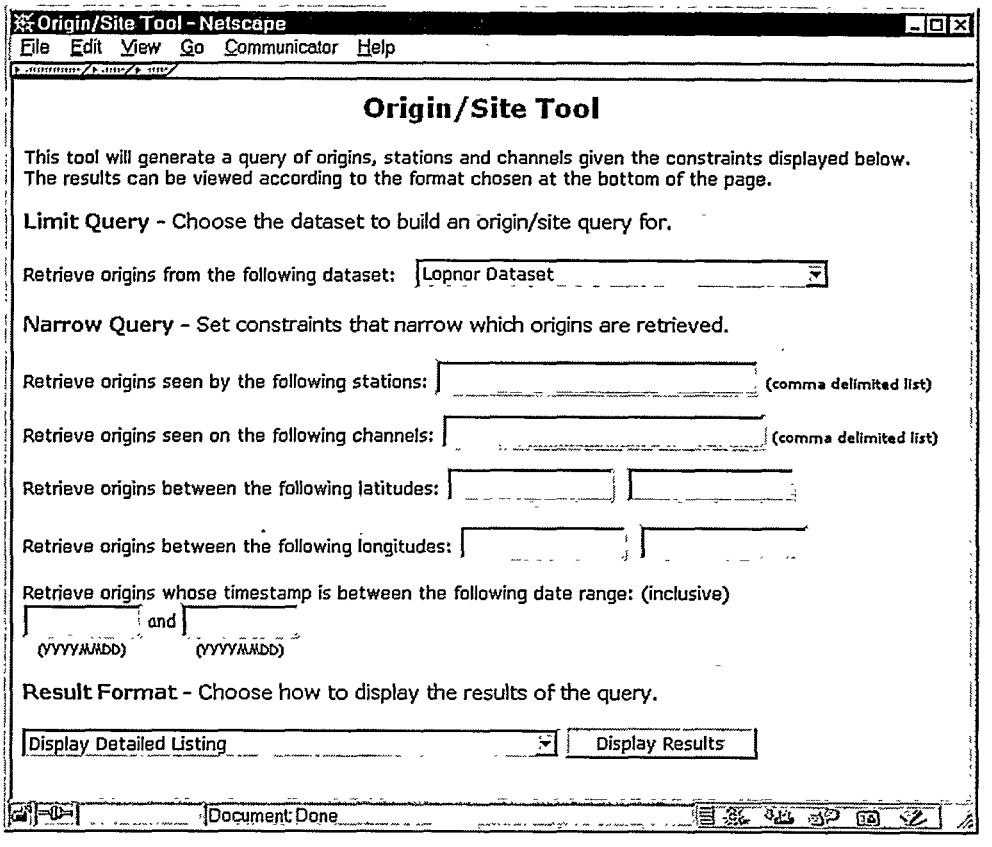


Figure 4. Content Analysis Tools Query Screen.

The results of these queries can be displayed in different user-selectable ways. The detailed listing for origins with waveforms produces a table containing the origins with their associated stations. In the corresponding table cells, waveform or arrival availability is noted by channel. The listing provides researchers with specific information about the origins that the charts don't include. The charts, however, include information on 1) total number of origins per station ranked from most to least origins seen, 2) total number of origins per station broken down by year when seen ranked from most to least total origins (See Figure 5), 3) total number of origins per year, 4) number of origins per number of stations, and 5) number of origins per number of stations broken down by year when seen. The charts are extremely useful to the researchers in presenting the "big picture" aspects for the data set being queried. These aspects include such things as the distribution of origins to stations, origin coverage, yearly coverage, etc.

Like the dendrogram tool, it helps to visualize the results based on the location of the origins on a map. Thus an ArcView based version of the query interface was built to complement the web-based applications. Using this

interface, the researcher can perform the exact same query they used for the web-based tools to present the resulting origins geo-located on a map. The origins are also color-coded based on the year of the event. This makes it easier to compare with the plots broken down by year. (See Figure 6)

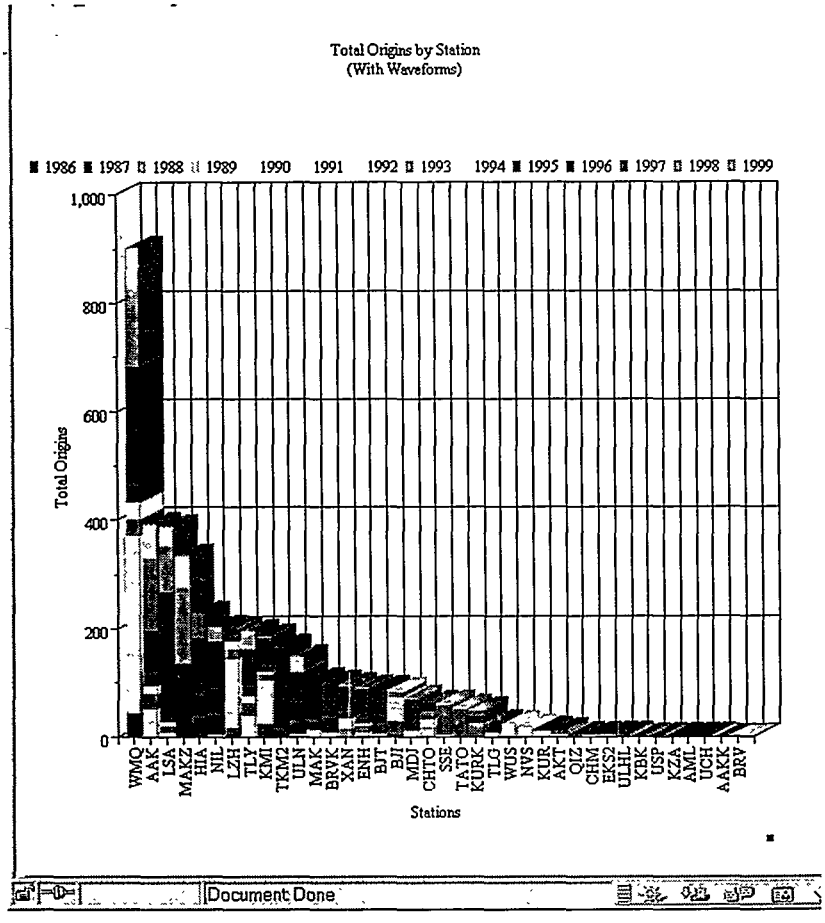


Figure 5. Sample Total Origin Plot by Station broken down by year.

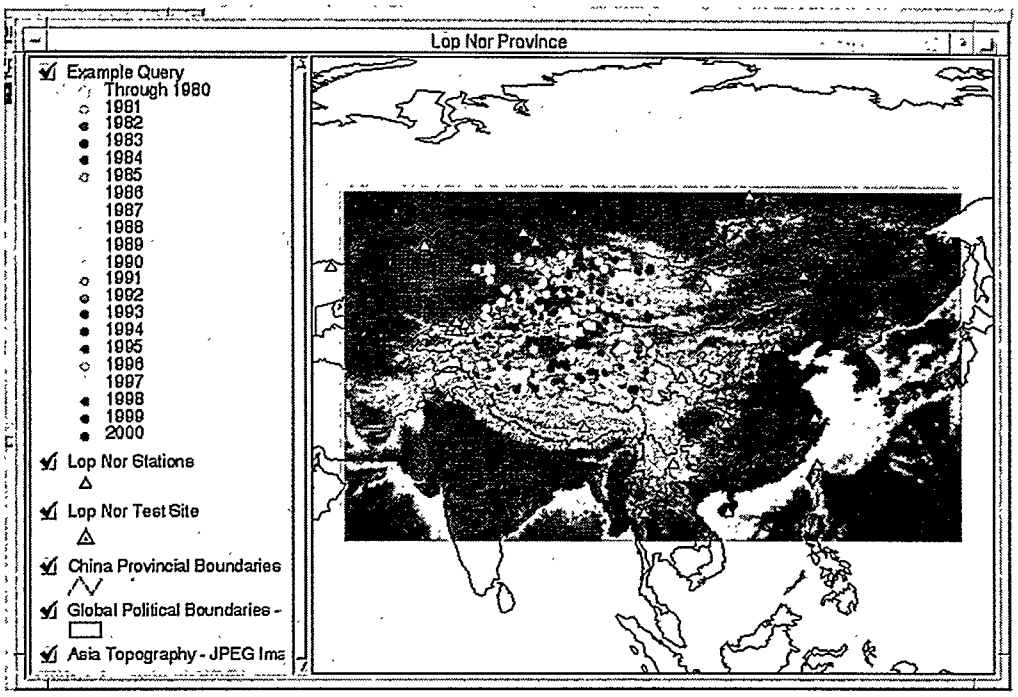


Figure 6. ArcView Sample Results Screen.

3) *WebDB*:

In the same way the web-based Content Analysis tools allow the researcher to query the database without having to know much, if anything, about SQL, Oracle's WebDB can be used by researchers to manage, explore, and manipulate seismic data in their research databases. WebDB provides an out-of-the-box web-based interface to look at the data contained within the database. All that is required from the researcher to use this tool is a web browser, such as Netscape® Navigator or Microsoft® Internet Explorer.

One of the major problems associated with storing seismic data in a sophisticated relational database management system is knowing how to efficiently access and use this data. Until now researchers had to learn SQL in order to access and display the data contained in a relational database system. WebDB is an off-the-shelf software application that gives researchers the ability to access reports that dynamically query the database tables containing seismic data using selection criteria specified by the researcher. The researcher can run a report and specify, for example, a latitude-longitude window and a time range and see the results displayed in either html, plain text ASCII, or MS Excel formats. Researchers with an intimate knowledge of the underlying database structure can create their own reports and make them available to other researchers. One of the useful features of WebDB is that these reports can contain links to other WebDB components, such as other reports with more detailed information and their associated charts, which can be used to graphically display the results of the pre-created query. The ability to display results in a graphical manner using WebDB is not as advanced as the graphical capabilities of the Content Analysis Tools described above; however, it provides charting options at a more basic level.

How can WebDB help a seismic researcher?

Following is a simple example of the practical application of using WebDB to look at the contents of a seismic research database.

A researcher accesses via a web browser a report that contains a summary list of events found in a given area of interest. The researcher creates this report by simply providing a latitude range and longitude range in a parameter entry form similar to the one shown in Figure 7.

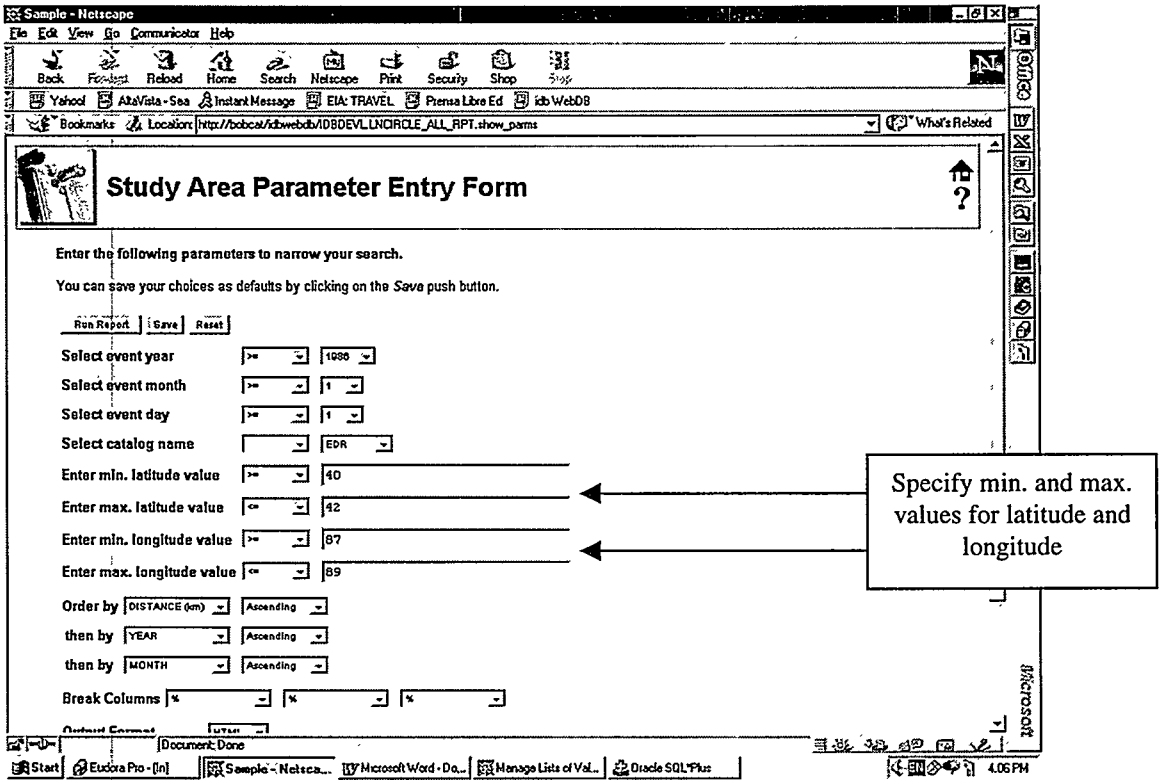
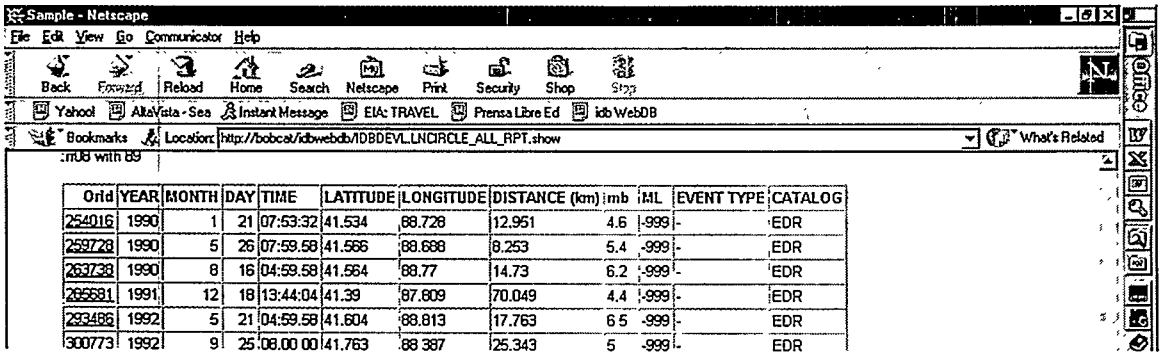


Figure 7. WebDB Create Report Sample Sceen.

This report contains event summary information such as the origin identification number (origin id), latitude, longitude, depth, magnitude, and time of the events. (See Figure 8)



Orid	YEAR	MONTH	DAY	TIME	LATITUDE	LONGITUDE	DISTANCE (km)	mb	ML	EVENT TYPE	CATALOG
254016	1990	1	21	07:53:32	41.534	88.728	12.951	4.6	-999	-	EDR
259728	1990	5	26	07:59.58	41.566	88.688	8.253	5.4	-999	-	EDR
263738	1990	8	16	04:59.58	41.564	88.77	14.73	6.2	-999	-	EDR
285681	1991	12	18	13:44:04	41.39	87.609	70.049	4.4	-999	-	EDR
293486	1992	5	21	04:59.58	41.604	88.813	17.763	6 5	-999	-	EDR
300773	1992	9	25	08.00 00	41.763	88 387	25.343	5	-999	-	EDR

Figure 8. WebDB Sample Report.

By clicking on the origin id link, the researcher will then be presented a report that contains a deeper level of detail regarding the chosen event, such as Julian date and epoch time, number of associated phases, number of locating phases, geographic region number, seismic region number, estimated depth from depth phases, and much more event-specific information. (See Figure 9)

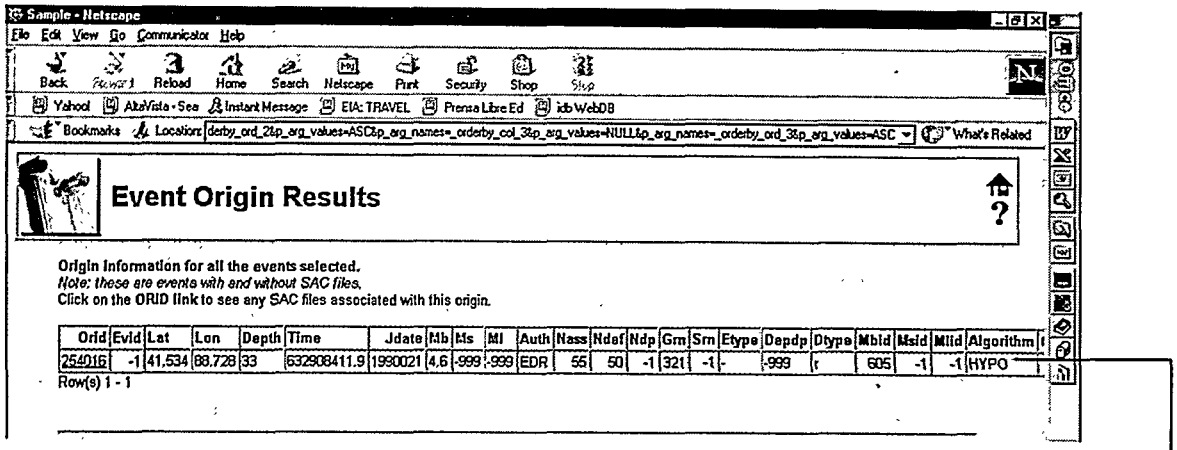


Figure 9. WebDB Linked Report Sample.

The researcher can then continue the drill-down process and, by clicking on the event link from this more-detailed report, he can see a third report which contains a listing of all the waveform files associated with the chosen event. (See Figure 10) This waveform file report can then be formatted and saved as a plain text ASCII file, which can in turn be used as data input to other applications, such as SAC (Seismic Analysis Code).

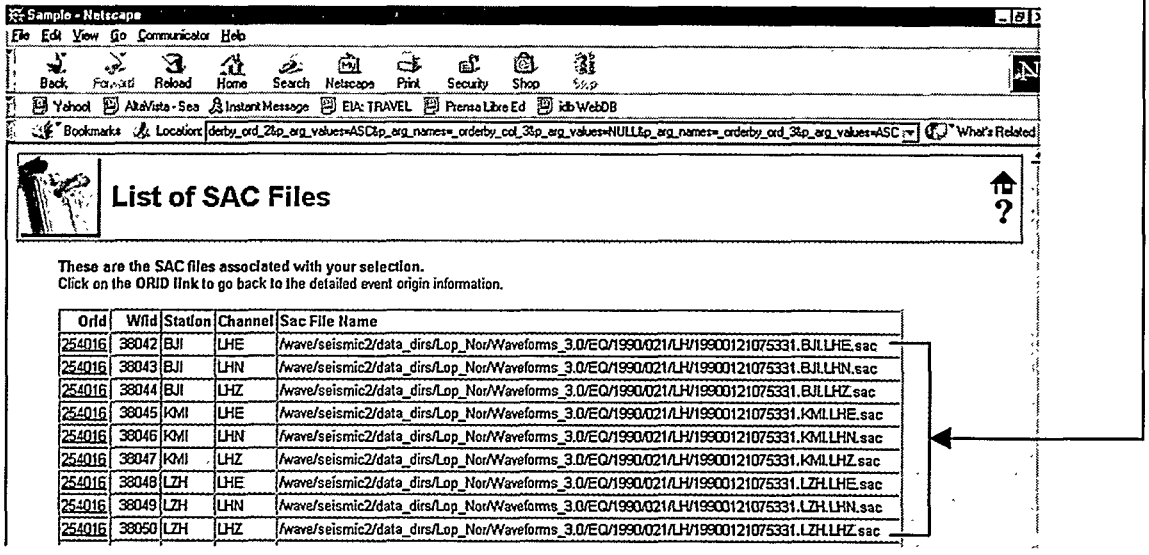


Figure 10. WebDB Drill-Down Example.

Alternatively, the origin id link can take the researcher to a different report which shows information pertaining to all the arrivals recorded for this event, such as station names that recorded the event, channels, phases recorded with those channels, arrival times, signal type, azimuth, amplitude of measurement, period, signal to noise ratio, and many more parameters.

CONCLUSIONS AND RECOMMENDATIONS

Each of the tools presented in this paper deal with different aspects of information overload. The Event Search Engine via the dendrogram tool can be used to find similar waveforms from which a possible reduced set can be generated. The content analysis tools help the researcher to ask "big picture" type questions about their data sets and present the results in various forms. Oracle's WebDB provides a simple wizard-style approach to creating interfaces for viewing, using, and managing relational database data.

For the Event Search Engine, our future work will focus on necessary modifications to process very large (>500) sets of waveforms. We believe that to do this, the data set must be broken down into subsets, which can then be processed sequentially. As each subset is processed, all redundant, sufficiently similar, waveforms are winnowed and the resulting set is passed on to process with the next subset. Because many of the waveforms may have been automatically collected and thus may not contain good signals, we will also need to build in a mechanism to weed out low quality waveforms which add nothing to the resultant dendrogram.

We believe that we have just scratched the surface on the kinds of listings and plots that can be provided by the content analysis tools. We think these other listings and plots will be even more useful in providing insight into a researcher's data set. However, even at this early stage of development, we have found these tools to be useful for various purposes. On the one hand, the researchers will find it a useful tool to find holes in their data sets where they might need to do more data gathering to build a more complete data set. While on the other hand, the researcher who wants to use a data set will find it a useful tool for browsing the data set content to get a feel for what is in the particular data set they are interested in. In the near future we would like to expand the plotting capability to multiple dimensions, such as plotting total origins by number of stations by magnitude, or total origins by station broken down by depth, etc. We will also be increasing the number of constraints by which a query can be limited by, such as station distance, azimuth gap, depth, magnitude, etc. All of these enhancements should combine to give the researcher tremendous power and accessibility to their database stored data sets.

REFERENCES

Aster, R. C. and J. Scott (1993). Comprehensive characterization of waveform similarity in microearthquake data sets, Bull. Seism. Soc. Amer., 83, 1307-1314.

Harris, D. B. (1991). A waveform correlation method for identifying quarry explosions, Bull. Seism. Soc. Amer., 81, 2395-2418.

Israelsson, H. (1990). Correlation of waveforms from closely spaced regional events, Bull. Seism. Soc. Amer., 80, 2177-2193.

Krzanowski, W. J. (1988). Principles of multivariate analysis: a user's perspective, Oxford Univ. Press, Oxford.

Ludwig, J. A., and J. F. Reynolds (1988). Statistical ecology, a primer on methods and computing, John Wiley & sons, New York, NY, 337 pp.

Riviere-Barbier, F. and L. T. Grant (1993). Identification and location of closely spaced mining events, Bull. Seism. Soc. Amer., 83, 1527-1546.