The Southern California Earthquake Data Center (SCEDC)

Katrin Hafner and Robert W. Clayton
Seismological Laboratory, California Institute of Technology

INTRODUCTION

The SCEDC is the primary archive facility for seismic information for southern California earthquakes. The data archive consists of earthquake parameters and travel-time picks from 1932 to the present, waveform recordings from 1981 to the present recorded by the Southern California Seismic Network (SCSN) and Caltech-USGS TriNet, SAR satellite images of southern California, and some regional refraction profiles of the crust and mantle. Data from portable instrument deployments after the 1992 Joshua Tree and Landers earthquakes, and from the 1994 Northridge earthquake, are also stored in this facility. The Data Center is currently archiving nearly 3,000 data channels from 375 stations. An average of 20,000 earthquakes are processed and archived each year.

The history of the SCEDC, the current state of data collection, the design of the database and archival system, user access to the system, and the plans for the future are outlined in this article. In particular, we discuss how the data from a large seismic array are stored, and we explain modern methods for accessing the data.

HISTORY OF THE SCEDC

The SCEDC facility was initiated in October 1991 as part of the Southern California Earthquake Center (SCEC). The several thousand computer tapes of the Caltech/USGS Seismic Processing (CUSP/SCSN) archive were translated into a custom ASCII database containing parametric earthquake data. The triggered seismograms for over 300,000 earthquakes were stored on an Internet-accessible 0.6 Tbyte optical WORM mass-storage system. The user access to this archive was initially through direct login to the Data Center machines. Parametric data (*e.g.*, hypocenters and phase picks) were later made available through a Web interface (http://www.scecdc.scec.org/catalog-search.html). Until late 1999, when TriNet, the new modern digital broadband array (Hauksson *et al.*, 2001), came online, the SCEDC archive consisted primarily of short-period, 100 sample/second waveforms from the SCSN. Since September 1999 the Data Center archive has contained waveforms from over 150 broadband and 200 accelerometer instruments, as well from the original SCSN short-period vertical stations. Data transfer between the monitoring network and the Data Center has also changed significantly since its inception. A time delay of a few days used to be the standard for new data to be available at the SCEDC. With the inception of the TriNet system and with changes in the daily operations and daily archiving of the Data Center, new earthquake data are available to the community in near real-time. The current operations, data archiving, and data access to the Data Center are described below.

CURRENT CONFIGURATION OF THE SCEDC

The SCEDC consists of two Sun 450 servers that handle the database software, the mass storage system, and temporary archives of waveform data. A separate Sun computer handles the Web server traffic for http://www.scecdc.scec.org, and user accounts are housed on another.

The mass-storage device for the SCEDC was upgraded in 1999 to a DISC-1050 magneto-optic disk system. The storage media are 5¼″ inch "WORM" or rewritable disks, with each disk holding 5.4 gigabytes of data. The total online storage capacity of this system is 5 terabytes. The mass-storage system is managed by SAM-FS archival software that makes the device appear as a large Unix file system. One of the valuable features of this software is that it allows high-demand data to be cached automatically on more rapidly accessed magnetic disks.

The hardware/software configuration of the SCEDC has been developed in conjunction with the Northern California Earthquake Data Center (NCEDC). Historically, these two groups have tried to develop similar systems so that data can be easily exchanged.

Storage of Metadata at the SCEDC

Since January 2001, metadata, such as earthquake hypocenters, magnitudes, amplitudes, and pointers to waveform archives, are stored in an *Oracle* (version 8.1.7) relational database. During the coming year the data in the original custom-made database will be transferred to the *Oracle* system as well. The schemas used by the SCEDC/TriNet system to hold these data were developed in conjunction with the NCEDC and the USGS Menlo Park and are available at

http://quake.geo.berkeley.edu/db/. These schemas consist of three general groups: (1) parametric, (2) waveform, and (3) instrument response/hardware information.

The parametric schema is made up of tables with information regarding events (earthquakes), including locations, magnitudes, phase picks, and peak amplitudes. The waveform schema tables contain a list of all seismograms archived in the mass storage, along with the information necessary to retrieve them. The waveform information includes the start and stop time for all seismograms, including the windows of the continuous recordings. This allows users to search for waveforms easily by specifying a station/channel/time index. Users can also quickly retrieve triggered waveforms associated with an earthquake because of an association table that links waveforms to earthquakes (evids). The instrument-response/hardware-tracking schemas were developed by the NCEDC. These tables track station histories/responses as a function of time and are used to generate the instrument response information utilized in the creation of SEED volumes.

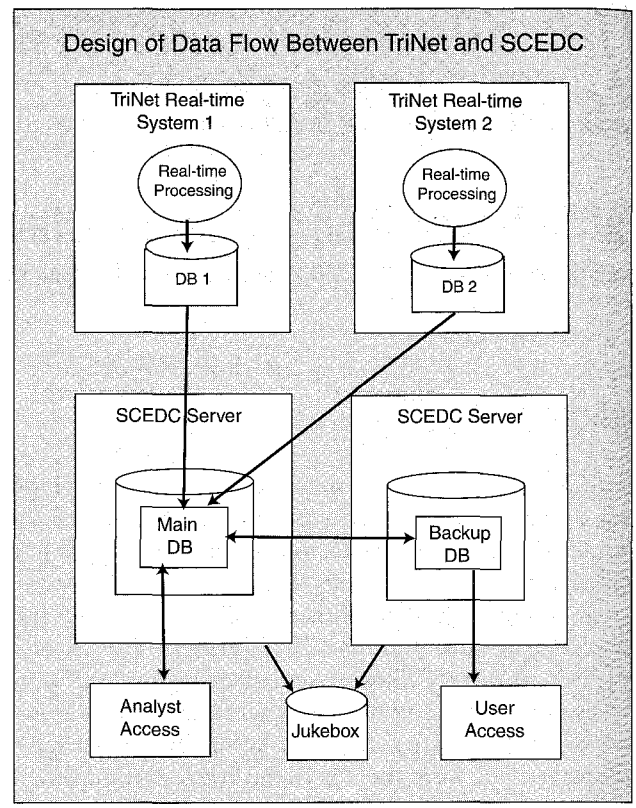
Two identical instances of *Oracle* databases are run by the SCEDC. These instances run on separate computers in order to provide backup in case of a computer failure and also to allow downtime of a database during periods of maintenance without interrupting user access. The data in the two databases are kept synchronized through two-way replication. Under normal operating conditions, the database load is divided between the two databases. TriNet analysts, as well as Data Center personnel, use one database that is protected from the outside world. Public users of the SCEDC archives are provided limited access to the other database. However, this configuration can easily be switched temporarily in the event that the normal operational situation has been disrupted.

Parametric Data Flow from TriNet to the SCEDC

In its present configuration, the SCEDC is tightly coupled to TriNet. The schematic view of this relationship is shown in Figure 1. Data that are generated by the TriNet real-time systems are written to the local databases that are resident on the real-time machines (Hauksson *et al.*, 2001). These data are then replicated to the "master" Data Center database. In the current configuration, there are two real-time systems reporting (what should be) identical data to the Data Center. These are distinguished from each other by having one of the real-time systems declared as "primary" and the other as "secondary." The data entries on the Data Center are then marked as coming from one or the other of the systems. The "primary" data are considered authoritative unless an error is detected, in which case they are replaced with the "secondary" data. This design allows for the complete failure of one of the real-time systems. The data in the real-time databases are periodically purged to keep the size of the real-time databases small and to maintain system performance. At least seven days of data are kept in the databases on the real-time systems.

Data from the TriNet real-time systems are normally "pushed" to the Data Center within a few seconds of their generation. If the Data Center machines are not available for some reason (*e.g.*, maintenance or system crash), then the real-time data are automatically queued up on the real-time databases. When the Data Center comes back online, the queued data are automatically transferred. This design was implemented for three main reasons. First, it allows the Data Center to go offline (intentionally or otherwise) without any loss of data. Second, the resynchronization occurs automatically and does not require human intervention. Finally, this design allows access to earthquake data on the real-time systems in an emergency should the Data Center be unavailable during an earthquake sequence. The real-time systems maintain all of the data recorded for at least the last seven days. The analysis tools that are normally used on the Data Center databases can be redirected to the real-time databases.

When an earthquake is detected by the real-time systems, basic parametric data and estimated hypocenter parameters are written into the database and are available for further analysis on the Data Center with a few seconds. These data are also immediately available to the general users of the SCEDC. Of course, users need to be aware that these data are in their raw form and could very well be in error. Over the course of the next day or so, analysts examine the data and make the necessary corrections and reestimation of the parameters.



▲ **Figure 1.** Diagram of Data Center: real-time system connection. The two real-time systems operate independently and populate their own local databases. These data are then replicated to one of the Data Center databases, which then replicates the data to the other Data Center database. The system can operate as a full-system with one real-time system and one data center machine if necessary, or with just one real-time system.

Waveform Data Flow between TriNet and the SCEDC

The data transfer of waveform data between the TriNet real-time systems and the Data Center is handled by a "pull" rather than "push" mechanism. The "pull" scheme ensures that the Data Center will not be overwhelmed by data being pushed to it to during a major earthquake sequence. When the real-time system processes an event, it also produces a set of time-window waveform requests ("request cards") for data to be archived. These are based on a magnitude-distance algorithm that also includes any seismograms that contributed to any of the "picked" data. A waveform retrieval program running on the Data Center begins the waveform retrieval process as soon as the local database receives these requests. These programs "pull" the data from wavepools on the real-time systems, which provide a complete reservoir of the past seven days of waveform data recorded by the network. The seven-day waveform reservoir gives the Data Center the flexibility of retrieving any requested waveforms over a period of seven days to avoid loss of data. This allows the Data Center to be offline for a considerable length of time. It also allows the network operators, or in fact anyone with permission, to "trigger" the network anytime up to seven days after the event time.

The retrieval of continuous waveform data is handled by a similar request card mechanism. In this case, the Data Center generates requests for 1-hour segments of data at the beginning of each hour. The continuous data are generally available within 15 minutes after the end of a particular 1-hour segment.

Both triggered and continuous waveform data are initially placed on a high-speed RAID system, where they can be rapidly accessed and easily augmented. After approximately seven days the data are automatically placed in a permanent archive form and stored on the mass-storage system. The seven-days duration is chosen because that is the maximum time for which any additional data can be requested from the wavepools on the real-time system. Where the data actually reside, *i.e.*, on the magnetic disk or in the jukebox archive, is transparent to the user.

SCEDC Data Archives

The archive of seismological data currently available at the Data Center is stored in different ways depending on the time period in question. Data from January 2001 to the present are archived in the *Oracle* database/archive system that is the focus of this article. The data acquired by the Data Center between circa 1980 and 31 December 2000 are stored in a custom database format and can be accessed by a set of tools that have been developed by the Data Center over the past decade, *e.g.*, *dbsort*, a catalog sorting program, and *scecgram*, a waveform retrieval program. The last group of data is the old archives (pre-1980 data and the hardcopy phase-pick archives), which require detailed human-intensive work to bring online. The plan is to convert the second type to the modern system over the coming year. The third type will be converted as time permits.

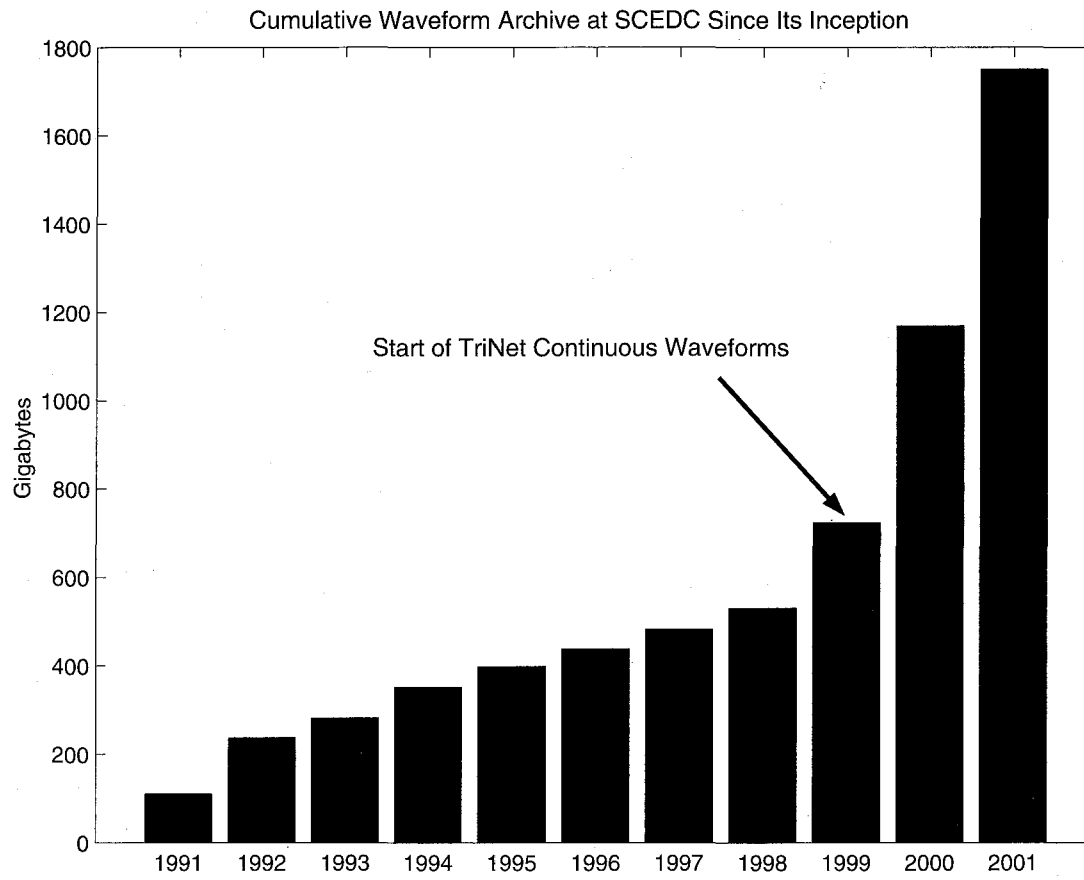
The current *Oracle* database is capable of storing extensive information about the earthquakes it records. There are provisions to archive the standard epicenter location parameters and their estimated errors, as well as mechanism and moment tensor solutions. While these values are not routinely populated for all events, the plan is to make these data available down to some minimum magnitude level. The database is also capable of storing multiple sets of parameters for an event. In particular, it can store the locations derived from the standard 1D Southern California velocity model, as well as those calculated with 3D velocity models such as Hauksson (2000) and the standard SCEC 3D velocity model (Magistrale *et al.*, 2000).

Post January-2001 Waveform Archives

The simplest approach to archiving waveform data is to store data as a continuous stream, as is done at the IRIS DMC. The penalty with this approach is the large mass-storage requirement, which would be 8.4 Tbytes/year for TriNet. The other end of the spectrum would be to store only triggered data that correspond to known events. Before TriNet, the Data Center archived only triggered waveform data. The downside to this approach is that important phenomena related to earthquake and other scientific studies may be missed.

To reduce the storage requirement for continuous archiving without sacrificing significant (or at least apparently significant) data, the SCEDC has adopted a mixed mode of archiving the data. In this scheme, all data with a sample rate of 40 sps or less are continuously archived. Data with a higher sample rate than this are archived only when they are part of a trigger. A trigger is any event (real or artificial) that causes all or part of the array to be recorded for a duration of a few seconds to a few minutes. Normally these triggers are local earthquakes declared by the real-time systems, but they can also be regional and teleseismic events declared by the NEIC. They can also be "artificial events" declared by the network staff, which forces the system to retrieve and archive waveforms for some interesting phenomena, such as shock waves generated when the space shuttle lands at Edwards Air Force Base. With the seven-day reservoir on the real-time systems, these triggers can be activated anytime up to seven days after an event. In the event of a major earthquake or swarm, the Data Center has the option of turning on continuous archiving of the entire network. Figure 2 illustrates the storage requirements for triggered and continuous waveform data since the Data Center's inception. This mixed triggered-continuous storage reduces the modern mass-storage requirements to about 0.6 Tbytes per year.

The data for a given trigger are archived in a few large files, in which the seismogram window segments are placed end to end with the location of a given seismogram within the file stored in the database. The retrieval of a suite of seismograms for any given event is very quick, since it usually only involves recovering one file from the archive. In order to avoid possible data translation errors, the data are archived in binary format (usually Mini-SEED in 512-byte blocks) as



▲ **Figure 2.** Waveform archives at the SCEDC in gigabytes. The 1991 data archive includes all of the waveform data from 1981 through 1991. The increase starting in 1999 is due to the archiving of continuous broadband data. The last four months of 2001 are estimated.

delivered by the real-time systems. They are converted automatically to other formats when a user retrieves the data.

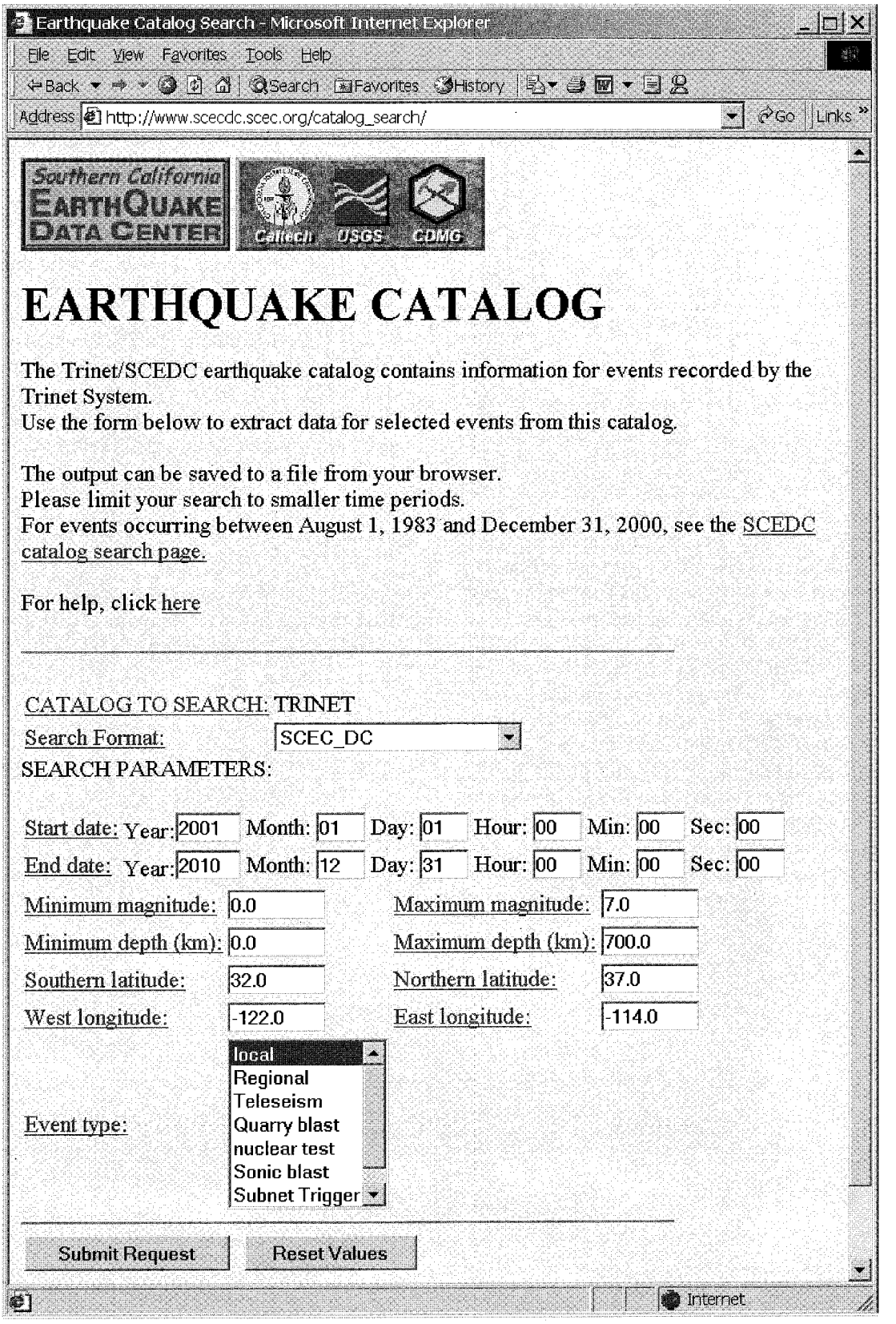
The continuous waveform data are retrieved in 1-hour time windows for the 40 sps data and 1-day segments for the lower sample rates. The 24 1-hour segments (*i.e.*, one day) of data are then archived in a single file, with each 1-hour segment being indexed in the database. This means that it is very efficient to retrieve seismograms on the order of an hour in length. The components of a particular data type (*e.g.*, BHZ, BHE, BHN) are also stored in the same file, since these are usually requested as a set.

USER ACCESS TO THE DATA

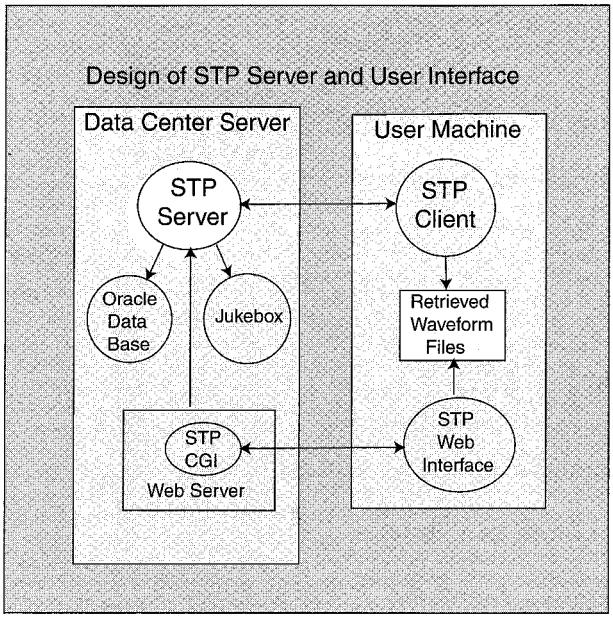
The traditional access to the Data Center archives for scientific users has been through direct login to the Data Center computers via individual user accounts. The program *dbsort* allows the user to sort the hypocenters, phase picks, and waveform data from 1981 through December 2000. *Scecgram* retrieves the data found by *dbsort* from the waveform archive. Scientific users can request an account on the Data Centers computers by logging on as "bulletin/bulletin" on the scec.gps.caltech.edu machine.

Access to the TriNet data (post-January 2001) is largely via various Web interfaces. The Web interface (http://www.scecdc.scec.org/catalog-search.html) allows the user to sort the January 2001 to present earthquake catalog (Figure 3). Over the last year, the SCEDC has developed a "seismic transfer program" called STP that allows direct transfer of the waveform data to a user's local disk (Figure 4). The Web-based version of this interface is available at http://www.scecdc.scec.org/stp.html (Figure 5), and a simple client version of this program is downloadable from http://www.scecdc.scec.org/software.html. With the Web-based version the user can search the catalog for waveform data. The selected waveforms are collected into a "tar" file, which can then be transferred to the user's computer via FTP. The user can retrieve both continuous and triggered data. Users can also develop preferred lists of stations that are retained between sessions in custom user profiles. The client version has the same functionality, but since it utilizes a direct network connection to the Data Center (a socket), the data are directly placed on the user's disk as they are retrieved. The data can be converted automatically to any of a number of formats, including SAC and Mini-SEED. The byte-swapping between machines of differing architectures is automatically taken care of. The client version is currently available as a simple C code for Unix and Linux systems.

For more traditional access to the continuous data, an IRIS BREQ_FAST type interface is provided to write full



▲ **Figure 3.** Web interface for the Earthquake Catalog search. This interface is the primary method for searching the parametric database. The URL is http://www.scedc.scec.org/catalogs.html.



▲ **Figure 4.** Diagram of the nature of the STP program. This program allows users to retrieve waveforms from the SCEDC archive and deliver them directly to the user's local machine. The STP program is accessible via a Web interface and a simple command-line client interface. Waveforms can be retrieved in user-defined formats such as SAC and Mini-SEED.

SEED volumes of the selected data. This method of data retrieval will be expanded to provide interoperability between other data centers by making a version of NETDC available.

We anticipate that all of the data archived at SCEDC will be available through the interactive STP and BREQ_FAST/NETDC interfaces in the coming year. We are also in the process of adding a mechanism for users to the waveform data interactively display and to select hypocenter and phase information graphically.

CONCLUSIONS

A database and data archiving scheme has been presented that provides the user immediate access to earthquake data for events in southern California and globally. The primary access for these data will be through interactive Web interfaces that deliver data directly to the user's computer.

ACKNOWLEDGMENTS

The SCEDC has received the bulk of its funding through the Southern California Earthquake Center (NSF EAR-8920136, USGS 14-08-0001-A0899, and 1434-HQ-97AG01718). In addition, direct funding (matched by funds from SCEC) for the current mass-storage system was provided by NSF (EAR 9816156). TriNet has also provided significant infrastructure support through support of database experts and computer hardware. Contribution number 300 from SCEC and 8818 from the Seismological Laboratory at Caltech.

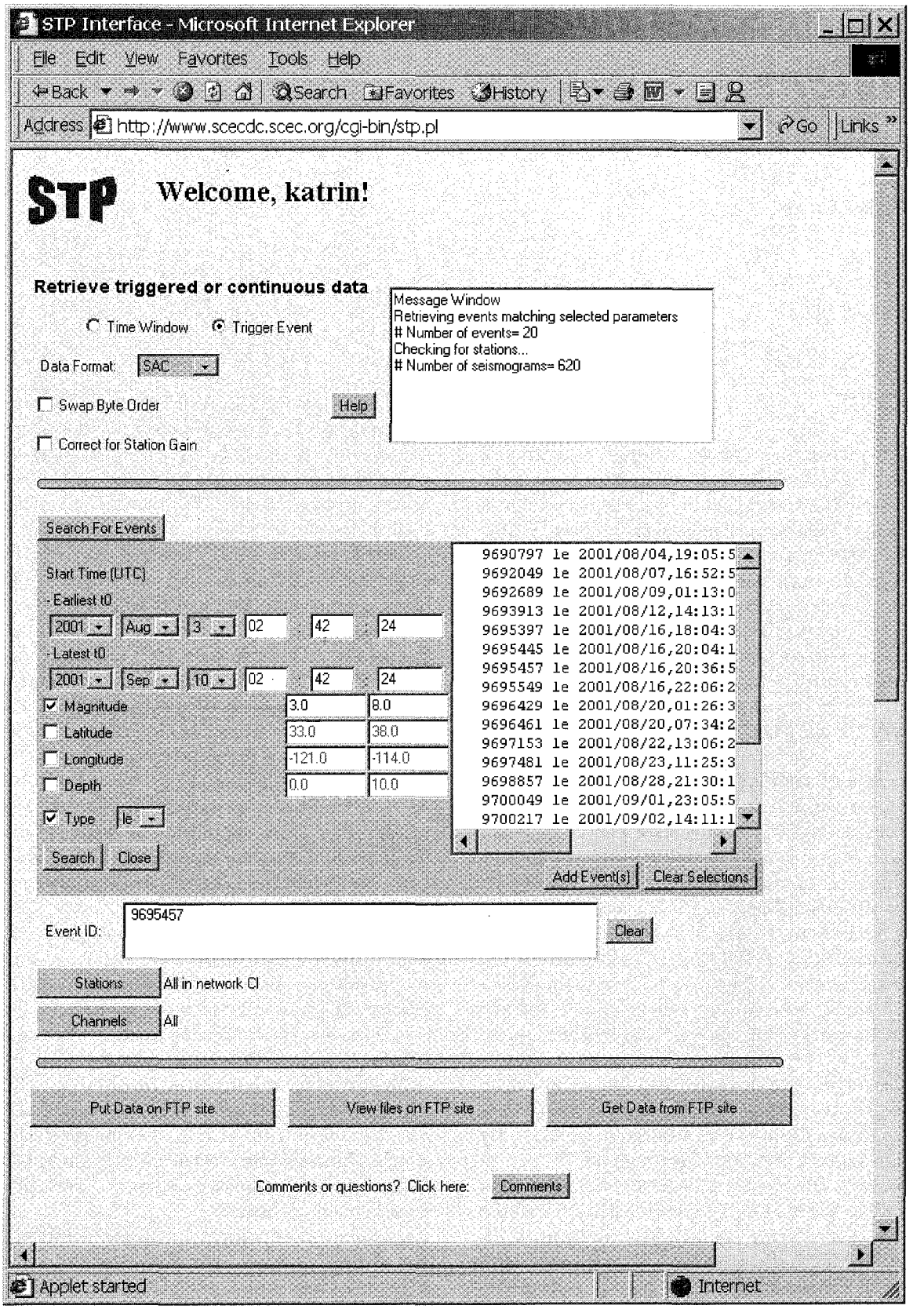
REFERENCES

Hauksson, E. (2000). Crustal structure and seismicity distribution adjacent to the Pacific and North America plate boundary in southern California, *J. Geophys. Res.* **105**, 13,875–13,093.

Hauksson, E., P. Small, K. Hafner, R. Busby, R. Clayton, J. Goltz, T. Heaton, K. Hutton, H. Kanamori, J. Polet, D. Given, L. Jones, and D. Wald (2001). Southern California Seismic Network: Caltech/USGS Element of TriNet 1997–2001, *Seism. Res. Lett.* **72**, 690–711.

Magistrale, H., S. Day, R. Clayton, and R. Graves (2000). The SCEC southern California reference three-dimensional model version 2, *Bull. Seism. Soc. Am.* **90**, S65–S76.

Seismological Laboratory
California Institute of Technology
Pasadena, CA 91125
clay@gps.caltech.edu



▲ **Figure 5.** STP Web interface with the Event Search option. This Web interface will retrieve waveform data from both the triggered archive and the continuous archive. The data are delivered in the format specified by the user. The URL is **http://www.scedc.scec.org/stp.html**.