

ANNUAL PROJECT SUMMARY

AWARD NUMBER: 98HQAG2219

Operation of the Northern California Earthquake Data Center

Barbara Romanowicz, P.I., and Doug Neuhauser
Berkeley Seismological Laboratory, UC Berkeley, CA 94720-4760
(510) 643-5690, x3-5811 (Fax), barbara@seismo.berkeley.edu
(510) 642-0931, x3-5811 (Fax), doug@seismo.berkeley.edu

PROGRAM ELEMENTS: I & II

KEY WORDS: Seismology; Geodesy; Database

INVESTIGATIONS UNDERTAKEN

The Northern California Earthquake Data Center, a joint project of the Berkeley Seismological Laboratory and the U.S. Geological Survey at Menlo Park, serves as an "on-line" archive for various types of digital data relating to earthquakes in central and northern California. The NCEDC is located at the Berkeley Seismological Laboratory, and has been accessible to users via the Internet since mid-1992.

The primary goal of the NCEDC is to provide a stable and permanent archival and distribution center of digital geophysical data for northern and central California such as seismic waveforms, electromagnetic data, GPS data, and earthquake parametric data. The principal networks contributing seismic data to the data center are the Berkeley Digital Seismic Network (BDSN) operated by the Seismological Laboratory, the Northern California Seismic Network (NCSN) operated by the USGS, and the Bay Area Regional Deformation (BARD) GPS network. The collection of NCSN digital waveforms date from 1984 to the present, the BDSN digital waveforms date from 1987 to the present, and the BARD GPS data date from 1993 to the present.

RESULTS

Datasets

The bulk of the data at the NCEDC consist of waveform and GPS data from northern California. The total size of the datasets archived at the NCEDC is shown in Table 1.

BDSN/NHFN Seismic Data

The archival of current BDSN and Northern Hayward Fault Network (NHFN) seismic data is an ongoing task. BDSN and NHFN data are telemetered from 30 seismic data loggers in real-time to the BSL, where they are written to disk files. Each day, an extraction process creates a daily archive by retrieving all continuous and event-triggered data for the previous day. The daily archive is run through quality control procedures to correct any timing errors, triggered data is reselected based on the REDI, NCSN, and UCB earthquake catalogs, and the resulting daily collection of data is archived at the NCEDC.

Data Type	MBytes
BDSN broadband, electric field, and magnetic field waveforms	713,977
NCSN event seismograms	276,233
GPS data (RINEX and raw data)	235,870
Calpine/Unocal Geysers region seismograms	38,658
Parkfield HRSN seismograms	77,319
Nevada broadband seismograms	47,827
USGS Low Frequency geophysical data	907
Misc data	37,178
Total size of archived data	1,427,969

Table 1: Volume of Data Archived at the NCEDC by Data Type

All of the data acquired from the BDSN and NHFN Quanterra data loggers are archived at the NCEDC. Most of the 16-bit BDSN digital broadband data from 1987-1991 have been converted to MiniSEED and are now online. However, there are still data from the 16-bit BDSN data acquisition systems from MHC, SAO, and PKD1 from late 1991 through mid-1992 that need to be converted to MiniSEED. Likewise, data acquired by portable 24-bit RefTek recorders before the installation of Quanterra data loggers at NHFN sites has not yet been converted to MiniSEED and archived at the NCEDC.

NCSN/SHFN Seismic Data

NCSN and SHFN waveform data are now being sent to the NCEDC via the Internet. The NCSN event waveform files are automatically transferred from the Menlo Park to the NCEDC as part of the routine analysis procedure by the USGS, and are automatically verified and archived by the NCEDC. Parametric information, such as event catalogs and phase readings from both the BDSN and NCSN are automatically updated on the NCEDC on an hourly basis.

A few corrupt NCSN event files were discovered at the NCEDC several years ago, and were eventually traced down to suspected flaws in the 12-inch WORM media and/or firmware problems on the Sony WDA-600 series jukeboxes used by the NCEDC. When we transcribed the data from the 12-inch WORM media to the current 5.25 inch magneto-optical media, we verified that all files were transcribed accurately. This year, using software developed at the NCEDC to detect possibly corrupt NCSN files, we identified 4704 possibly corrupted NCSN waveform event files. We re-read the original NCSN tapes for all of these events, discovered that only 71 of the files were actually corrupt, and replaced the corrupted event waveform files.

The NCEDC maintains a list of teleseismic events recorded by the NCSN, which is updated automatically whenever a new NCSN event file is received at the NCEDC, since these events do not appear in the NCSN catalog.

Electro-Magnetic Data

The NCEDC continues to archive and process electric and magnetic field data acquired from data loggers at two sites (SAO and PKD). At PKD and SAO, 3 components of magnetic field and 2

or 4 components of electric field are digitized and telemetered in real-time along with seismic data to the Seismological Laboratory, where they are processed and archived at the NCEDC in a similar fashion to the seismic data. The system generates continuous data channels at 40 Hz, 1 Hz, and .1 Hz for each component of data. All of these data are archived and remain available online at the NCEDC. Using programs developed by Dr. Martin Fullerkrug at the Stanford University STAR Laboratory (now at the Institute for Meteorology and Geophysics at the University of Frankfurt), the NCEDC is computing and archiving magnetic activity and Schumann resonance analysis using the 40 Hz data from this dataset. The magnetic activity and Schumann resonance data can be accessed from the Web.

In addition to the electro-magnetic data from PKD and SAO, the NCEDC archives data from a low-frequency, long-baseline electric field project operated by Dr. Steve Park of UC Riverside at site PKD2. This experiment uses an 8-channel Quanterra data logger to record the data, which are transmitted to the BSL using the same circuit as the BDSN seismic data. These data is acquired and archived in an identical manner to the other electric field data at the NCEDC.

Parkfield High Resolution Seismic Network Data

Event seismograms from the Parkfield High Resolution Seismic Network (HRSN) from 1987 through June 1998 are available in their raw SEG-Y format via NCEDC research accounts. A number of events have faulty timing due to the lack or failure of a precision timesource for the network. Due to funding limitations, there is currently no ongoing work to correct the timing problems in the older events or to create MiniSEED volumes for these events. However, a preliminary catalog for a significant number of these events has been constructed, and the catalog is available via the web at the NCEDC.

The original HRSN acquisition system died in late 1998, and an interim system of portable RefTek recorders was installed at some of the sites. Over the past year, 3 new borehole sites were installed, the system has been upgraded to operate with Quanterra Q733 data loggers and digital telemetry to new central processing site in Parkfield. Event files from the HRSN network are automatically transferred to the NCSN, and are made available to the research community via anonymous ftp until they are reviewed and permanently archived.

As an interim measure, continuous data from the HRSN has been temporarily archived on the NCEDC in order to help researchers retrieve events that were not detected during the network upgrade.

GPS Data

The NCEDC continues to expand its archive of GPS data from the BARD (Bay Area Regional Deformation) network of continuously monitored GPS receivers in northern California. The NCEDC GPS archive now includes 40 continuous sites in northern California. There are 24 core BARD sites owned and operated by UC Berkeley, LLNL, USGS, UC Davis, Trimble Navigation, and Stanford. Data from the other northern California sites are collected from sites operated by JPL, the U.S. Coast Guard, and Scripps Institute of Oceanography.

Most of the Berkeley BARD sites are collocated with seismic stations, and data from these sites are acquired in real-time using shared frame relay telemetry link. The remaining Berkeley BARD stations use dedicated frame relay and/or spread spectrum radio to provide data in real-time to UC Berkeley, and are automatically processed and archived at the NCEDC on a daily basis. Data from

the USGS sites are downloaded by the USGS and transferred to the NCEDC on a daily basis, and is automatically archived by the NCEDC. The other sites are automatically acquired from their respective operators on an hourly or daily basis, and are archived by the NCEDC.

This year the NCEDC continued to archive non-continuous survey GPS data. The initial dataset to be archived is the survey GPS data collected by the USGS Menlo Park for northern California and other locations. The NCEDC will be the principal archive for this dataset. Significant quality control efforts were implemented by the NCEDC to ensure that the raw data, scanned site log sheets, and RINEX data are archived for each survey. All of the USGS Menlo Park GPS data has been transferred to the NCEDC and the majority of the data has been archived and is available for distribution.

Calpine/Unocal Geysers Seismic Data

The Calpine Geysers seismic network, consisting of up to 49 different borehole monitoring sites, was initially deployed and operated by the Unocal Geothermal Division on behalf of the geothermal energy producers in the California Geysers geothermal field. In 1999, the Unocal geothermal fields at the Geysers were acquired by the Calpine Corporation. Calpine has continued Unocal's collaboration with the NCEDC and has made data available for public distribution. The Calpine/Unocal Geysers dataset is a collection of digital microearthquake seismograms recorded by this network. 12 years of digital microearthquake seismograms (1987 through 1998) have been released for archiving and distribution through the NCEDC.

Calpine has also released an unedited earthquake catalog for events in 1995-1998 with approximate times and very rough hypocenters for some of the event represented in the time series, primarily as an index aid for the waveforms. Calpine makes no claims that this catalog is complete in any manner, and assumes no responsibility for the accuracy or usefulness of the catalog data.

Through an updated agreement with the NCEDC last year, Calpine released triggered event waveform data and a preliminary hypocenter catalog for an additional two years of data from 1999-2000. The total dataset represents over 248,000 events that were recorded by the Calpine/Unocal Geysers network, and is available via research accounts at the NCEDC.

USGS Low Frequency Data

Over the last 25 years, the USGS at Menlo Park, in collaboration with other principal investigators, has collected an extensive low-frequency geophysical data set that contains over 1300 channels of tilt, tensor strain, dilatational strain, creep, magnetic field, water level, and auxiliary channels such as temperature, pore pressure, rain and snow accumulation, and wind speed. We are actively working with the USGS to assemble the requisite information for the hardware representation of the stations and the instrument responses for this diverse dataset.

During the past year, we installed the necessary programs to archive the raw data in standard MiniSEED and to populate the database with the necessary parameters to create SEED instrument responses. We have currently archived timeseries data from 887 data channels from 167 sites, and have instrument response information for 542 channels at 139 sites. The waveform archive is updated on a daily basis with data from 486 currently operating data channels.

There is also considerable interest in having the NCEDC archive and distribute a "processed" version of the principal data channels from this data set. We will augment the raw data archive as

additional instrument response information is assembled for the channels, and will work with the USGS to clearly define the attributes of the "processed" data channels.

Northern California Seismicity Project

The objective of the Northern California Seismicity Project (NCSP), which commenced this year, is to transcribe the pre-1984 data for $M_L \geq 2.8$ earthquakes which have occurred in Northern and Central California (NCC) outside of the San Francisco Bay region (SFBR), from the original reading/analysis sheets of the Berkeley Seismological Archives, into a computer readable format. This work complements the ongoing Historical Earthquake Relocation Project (HERP) of the Berkeley Seismological Laboratory, which concentrates solely on San Francisco Bay Region.

The characterization of the spatial and temporal evolution of NCSP seismicity during the initial part of the earthquake cycle as the region emerges from the stress shadow of the great 1906 San Francisco earthquake is the long term goal. The problem is that the existing BSL seismicity catalog for the SFBR, which spans most of the past century (1910-present), is inherently inhomogeneous because the location and magnitude determination methodologies have changed, as seismic instrumentation and computational capabilities have improved over time. As a result, NCC seismicity since 1906 is poorly understood.

Creation of a NCC seismicity catalog that is homogeneous, that spans as many years as possible, and that includes formal estimates of the parameters and their uncertainty is a fundamental prerequisite for probabilistic studies of the NCC seismicity. The existence of the invaluable BSL seismological archive, containing the original seismograms as well as the original reading/analysis sheets, coupled with the recently acquired BSL capability to scan and digitize historical seismograms at high resolution allows the application of modern analytical algorithms towards the problem of determining the source parameters of the historical SFBR earthquakes.

The funding level for this project did not allow us to transcribe all of the pre-1984 reading/analysis sheets from the Berkeley Seismological Archive. However, limiting our work to earthquakes of $M_L \geq 2.8$ provides a significant contribution to the uniformity of the NCC seismicity catalog. We anticipate continuing with this project as funding allows.

CNSS Catalog

The NCEDC, in conjunction with the Council of the National Seismic System (CNSS), is producing and distributing a world-wide composite catalog of earthquakes based on the catalogs of the national and various U.S. regional networks. Each network updates their earthquake catalog on a daily basis at the NCEDC, and the NCEDC constructs a composite world-wide earthquake catalog by combining the data, removing duplicate entries that may occur from multiple networks recording an event, and giving priority to the data from each network's *authoritative region*. The catalog, which includes data from 14 regional and national networks, is searchable using a Web interface at the NCEDC. The catalog is also freely available to anyone via ftp over the Internet.

Hardware and Software system

The NCEDC is housed with the computing facilities at the Berkeley Seismological Laboratory in McCone Hall. The BSL and NCEDC computers share a high-speed 100 MBit switched network.

In 1998 partial funding for a new mass storage system was made available from the USGS in mid-year, but the purchase of the new mass storage system had been deferred in an attempt to acquire higher density drives, which ultimately were not available. The new mass system was purchased in late spring 1998, and is comprised of a Sun Ultra 450 computer, a 1.3 Tbyte DISC 517 slot jukebox with two 2.6 GByte Magneto Optical (MO) drives, an 11-slot AIT tape jukebox which holds 25 GBytes per tape, and the SAM-FS hierarchical filesystem management software. Only two MO drives and minimal media were purchased at that time.

In 1999, the mass storage system was upgraded from its initial 1.3 TByte capacity to 2.5 TByte capacity by the replacement of its 2.6 GByte MO drives by four 5.2 GByte MO drives and 5.2 GB MO media. The mass storage system can be upgraded to a total of 1000 slots (5.2 TByte) capacity with the addition of a second media picker, drives, and media cells.

The new hardware and software system can be configured to automatically create multiple copies of each data file. The NCEDC is using this feature to create an online copy of each data file on MO media, and another copy on AIT tape which will be stored offline.

During the past year, the NCEDC updated memory in its two Sun servers.

Joint Northern California Earthquake Catalog

Currently both the USGS and BSL construct and maintain earthquake catalogs for northern and central California. The "official" UC Berkeley earthquake catalog begins in 1910, and the USGS "official" catalog begins in 1966. Both of these catalogs are archived and available through the NCEDC, but the existence of 2 catalogs has caused confusion among both researchers and the public. The BSL and the USGS have spent considerable effort over the past year to define procedures for merging the data from the two catalogs into a single northern and central California earthquake catalog in order to present a unified view of northern California seismicity. The differences in time period, variations in data availability, and mismatches in regions of coverage all complicate the task.

From 1910 through 1967, the BSL catalog is the primary source of northern California earthquake information. Only limited phase data are available for this time period, although location and magnitude information is provided. The NCSN began to come online in 1966, and observations from this network are available beginning in July of that year.

Starting with data from 1996, the BSL and USGS are working to generate a "joint" catalog by merging phase data and relocating the earthquakes. One of the initial complications in this project is matching up events between the two catalogs. Due to the sparse nature of the BSL instrumentation over the years, the BSL catalog is only complete at the magnitude 3 level while the USGS catalog is generally complete at the magnitude 2 level. However, the BSL catalog includes regional events from southern California, Nevada, Utah, Oregon, and Washington. Thus neither catalog is a subset of the other. Other complications include foreshock/aftershock sequences, where one organization might read a foreshock and the mainshock and the other might read the mainshock and an immediate aftershock. Since limited phase data are available for the BDSN until 1976, most events during this period will combine the USGS location with the BSL magnitude. Where BSL phase data are available, an event that appears in both catalogs will have its phase and amplitude data merged, the event will be relocated with the combined phase data, and magnitudes will be recomputed using the available amplitude readings and new location.

The process of consolidating the BSL data to be merged into the joint catalog has uncovered

many details in the original catalog which were ambiguous or poorly documented or inconsistent over time, such as the use of channel names for phase and amplitude readings. This year, new problems over uncertainties in NCSN timing for older events have arisen.

The USGS and BSL performed an initial joint catalog from the USGS and BSL catalogs in the spring of 1999. The BSL spent considerable time analyzing the resulting catalog, and has identified problems with specific earthquake associations and other related problems. Most of the remaining problems are associated with events outside the network, especially in the Cape Mendocino/Gorda plate region. We are working to finalize this effort as well as to establish procedures to creating an on-going joint catalog.

User Interface Development

During the past year, the NCEDC has developed a generalized database query system to support the development of portable database query applications among data centers with different internal database schemas. The initial goal was to modify the IRIS SeismiQuery web interface program to make installation easier at the NCEDC and other data centers, as well as to introduce a new query language that would be schema independent.

In order to support SeismiQuery and other future database query applications, we defined a set of Generic Data Views (GDV) for the database that encompassed the basic objects that we expect most data centers to support. We introduced a new language we call MSQL (Meta SeismiQuery Language), which is based on generic SQL, and uses the GDV's for its core schema. MSQL queries are converted to Data Center specific SQL queries by the parsing program MSQL2SQL. This parser stores the MSQL parsing tree in a data structure and API's were implemented to browse and modify elements in the parsing tree. These API's are the only datacenter or database specific source codes. We finally modified the SeismiQuery web interface to uniformly generate MSQL requests and to process these requests in a consistent fashion.

We have installed SeismiQuery at the NCEDC, where it provides a common interface for querying attributes and available data for both the BDSN and the USGS Low Frequency networks.

We have provided both IRIS and the SCEC Data Center with our modified version of SeismiQuery. We envision using this approach to support other database query programs in the future.

Database Development

Most of the parametric data archived at the NCEDC, such as earthquake catalogs, phase and amplitude readings, waveform inventory, and instrument responses have been stored in flat text files. Flat file are easily stored and viewed, but are not efficiently searched. Over the last year, the NCEDC, in collaboration with the USGS/SCEC Data Center, and TriNet, has continued development of database schemas to store the parametric data from the joint earthquake catalog, station history, complete instrument response for all data channels, and waveform inventory.

The parametric schema supports tables and associations for the joint earthquake catalog. It allows for multiple hypocenters per event, multiple magnitudes per hypocenter, and association of phases and amplitudes with multiple versions of hypocenters and magnitudes respectively. The instrument response schema represents full multi-stage instrument responses (including filter coefficients) for the broadband data loggers. The hardware tracking schema will represent the inter-connection of instruments, amplifiers, filters, and data loggers over time. This schema will be used to store the joint northern California earthquake catalog and the CNSS composite catalog.

The entire description for the BDSN and USGS Low Frequency Geophysical networks and data archive has been entered into the hardware tracking, SEED instrument response, and waveform tables. Programs have been developed to perform queries of waveform inventory and instrument responses, and the NCEDC can now generate full SEED volumes from the BDSN network based on information from the database and the waveforms on the mass storage system. The second stage of development will include the NCSN waveform inventory and later the NCSN instrument response data as they are made available. We distributed all of our programs and procedures to populate the hardware tracking and instrument response tables to the SCECDC in order to help them populate their database.

Additional details on the joint catalog effort and database schema development may be found at <http://quake.geo.berkeley.edu/db>

Data Distribution

The NCEDC continues to use the World Wide Web as a principal interface for users to request, search, and receive data from the NCEDC. The NCEDC has implemented a number of useful and original mechanisms of data search and retrieval using the World Wide Web, which are available to anyone on the Internet. All of the documentation about the NCEDC, including the research users' guide, is available via the Web. Users can perform catalog searches and retrieve hypocentral information and phase readings from the various earthquake catalogs at the NCEDC via easy-to-use forms on the Web. In addition, users can peruse the index of available broadband data at the NCEDC, and can request and retrieve broadband data in standard SEED format via the Web. Access to all datasets is available via research accounts at the NCEDC. The NCEDC's home page address is <http://quake.geo.berkeley.edu/>

In a collaborative project with the IRIS DMC and other worldwide datacenters, the NCEDC has helped develop and implement NETDC, a protocol which will provide a seamless user interface to multiple datacenters for geophysical network and station inventory, instrument responses, and data retrieval requests. The NETDC builds upon the foundation and concepts of the IRIS BREQ_FAST data request system. The NETDC system was put into production in January 2000, and is currently operational at three datacenters worldwide – the NCEDC, IRIS DMC, and Geoscope. The NETDC system receives user requests via email, automatically routes the appropriate portion of the requests to the appropriate datacenter, optionally aggregates the responses from the various datacenters, and delivers the data (or ftp pointers to the data) to the users via email.

The NCEDC is participating in the UNAVCO-sponsored GPS Seamless Archive Centers (GSAC) initiative, which is developing common protocols and interfaces for the exchange and distribution of continuous and survey-mode GPS data. During this year, the NCEDC developed and implemented procedures to generate the appropriate GSAC inventory records for previously archived GPS data at the NCEDC, and now automatically generates the GSAC inventory records on a routine basis as part of its archiving procedures. The GSAC inventory records are available via anonymous ftp for other GSAC data wholesalers and retailers.

The NCEDC hosts a web page that allows users to easily query the NCEDC waveform inventory, generate and submit NETDC requests to the NCEDC. The NCEDC currently supports both the BREQ_FAST and NETDC request formats. As part of our collaboration with SCECDC, the NCEDC provided its BREQ_FAST interface code to SCECDC, have worked closely with them to implement BREQ_FAST requests at the SCECDC.

The various earthquake catalogs, phase, and earthquake mechanism can be searched using NCEDC web interfaces that allow users to select the catalog, attributes such as geographical region, time and magnitude. The GPS data is available to all users via anonymous ftp. Research accounts are available to any qualified researcher who needs access to the other datasets that currently are not available via the Web.

NON-TECHNICAL SUMMARY

The Northern California Earthquake Data Center (NCEDC) is an on-line archive and distribution facility for waveform and catalog data for several regional networks: The Northern California Seismic Network (NCSN) operated by the U.S. Geological Survey (USGS), the Berkeley Digital Seismic Network (BDSN) and Parkfield High Resolution Seismic Network (HRSN) operated by the U.C. Berkeley Seismological Laboratory, the USGS Low Frequency Geophysical Data set, the Bay Area Deformation Array (BARD) operated jointly by U.C. Berkeley, USGS, and several other San Francisco Bay Area institutions. These data serve as basis for many research projects relevant to NEHRP goals in the Bay Area and central and northern California.

MEETING PRESENTATIONS

Neuhauser, D., D. Oppenheimer, S. Zuzlewski, L. Gee, M. Murray, W. Prescott, B. Romanowicz, New Projects and Datasets at the Northern California Earthquake Data Center, *Eos, Transactions, American Geophysical Union*, 81, 48, 2000.

DATA AVAILABILITY

Data are available from the Northern California Earthquake Data Center via the Internet at <http://quake.geo.berkeley.edu>. For additional information, contact Douglas Neuhauser at 510-642-0931 or doug@seismo.berkeley.edu.