Access and Visualization of Large Distributed Regional-Network Seismogram Data Sets

John N. Louie, Nevada Seismological Lab.; and Sushil J. Louis, Dept. of Computer Science The University of Nevada, Reno

in response to NSF 99-105, Large Scientific and Software Data Set Visualization

Project Description

Project Description Contents Page in Section Results From Related Prior NSF Support 1 2 Seismological Motivation Recent Results of Seismic Network Data Access and Visualization 4 Expected Results of Regional-Network Seismogram Access and Visualization 6 8 **Proposed Capabilities** Work Plan 14 PI Experience With Visualization of Distributed Data 14 Letter of Interest from the New Zealand National Seismograph Network 15

Results From Related Prior NSF Support of John N. Louie:

"Reflectivity Structure below the San Fernando Valley From Northridge aftershock recordings," EAR-9416224 \$46,978 for the period 1 Sept. 1994 through 28 Feb. 1996.

The main purpose of this completed project was to depict the reflectivity structure of a blind thrust fault system using earthquake data. We show that the aftershock recordings of the 17 January 1994 Northridge, California, earthquake have the spatial sampling needed to image local structures at a resolution comparable to the average distance between the events. Regional seismic networks in California that record digital seismograms from hundreds of stations make this crustal reflectivity profiling possible even in the absence of conventional active-source seismic data. Where our aftershock-source and LARSE active-source sections coincide in the San Gabriel Mts., both techniques show the same prominent mid-crustal bright spot (Chavez-Perez and Louie, 1998).

We imaged crustal reflective structures in the area below the Jan. 17, 1994 Northridge earthquake, using data from Northridge aftershocks recorded on the Southern California Seismic Network; and also below the 1991 Sierra Madre earthquake and its aftershocks. The seismograms are archived at the Southern California Earthquake Center (SCEC) Data Center in CUSP format. We developed the translation routines needed to put them into a format similar to that of an exploration-seismic data set, assisting us in visualizing, selecting, and processing the data.

Our work with data from aftershocks of the 1994 Northridge earthquake shows that we image crustal fault geometry, including the blind thrust that generated the 1994 event. The images test the existence and configuration of thrust ramps and detachments proposed from balanced-section reconstructions of shallow-crustal profiles and borehole data (Chavez-Perez and Louie, 1998). Our location of a 45° north-dipping reflector below the northern margin of San Fernando Valley, that cuts the Moho, supports thick-skinned rather than thin-skinned compressional tectonic models.

Education and Human Resources - This project supported Ph.D. research in part by UNR graduate students Sathish K. Pullammanappallil, A. M. Asad, and Sergio Chavez-Perez. Dr. Pullammanappallil received the Ph.D. in December 1994; and Dr. Chavez-Perez received the Ph.D. in December 1997. Dr. Chavez-Perez held a Fulbright Scholarship from Mexico that required research and tuition support in addition to his Fulbright stipend. This project was also instrumental in contributing toward the research computing maintenance and administration pool in the UNR Mackay School of Mines. Research contributions to this pool foster a high level of computing availability for undergraduate and graduate student research throughout the school.

Publications supported by related NSF contract to Louie:

- Chavez-Perez, Sergio, 1997, Enhanced imaging of fault zones in southern California from seismic reflection studies: Doctor of Philosophy Thesis in Geophysics at the Univ. of Nevada, Reno.
- Chavez-Perez, S., and J. N. Louie, 1995, Seismic reflection views of a blind thrust fault system using earthquake data: *Expanded Abstracts*, Soc. Explor. Geophys. 65th Ann. Internat. Mtg., October 8-13, Houston, Texas, p. 511-514.

- Chavez-Perez, S., and J. N. Louie, 1996, Reflectivity structure beneath San Fernando Valley, California, using seismicity data and exploration seismology techniques (expanded abstract): *Proc. Eleventh World Conf. on Earthquake Engin.*, Paper No. 907, Elsevier Science Ltd., June 23-28, Acapulco, Mexico.
- Chavez-Perez, S., and J. N. Louie, 1997, Isotropic scattering and seismic imaging of crustal fault zones using earthquakes: *Expanded Abstracts*, Soc. Explor. Geophys. 67th Ann. Internat. Mtg., Nov. 2-7, Dallas, Texas.
- Chavez-Perez, S., and J. N. Louie, 1998, Crustal imaging in southern California using earthquake sequences: *Tectonophysics*, 286 (March 15), p. 223-236.

Results From Related Prior NSF Support of Sushil J. Louis:

Grant number: 9624130 (1996 - 2000): CAREER: Combining Genetic Algorithms with Case-Based Systems

We had proposed to investigate systems that combine genetic algorithms with case-based reasoning systems. Such novel genetic-based machine learning systems increase performance at related tasks as they gain experience. We have established the feasibility of such systems and developed software implementations (in C and C++) as well as visualization tools (in Java) that allow multiple users on the web to run and view applications and their progress using nothing more than web browsers. The system and visualization interface codes are available from http://gaslab.cs.unr.edu/. We are now applying our research results to a variety of applications in seismology, vision, physics, engineering, and operations research. Selected recent publications acknowledging NSF support resulting from this work include:

- "Robust Stability Analysis of Discrete-Time Systems Using Genetic Algorithms," M. Sami Fadali, Sushil J. Louis, and Yongmian Zhang. IEEE Transactions on Systems Man and Cybernetics, Part B: Cybernetics, IEEE, to appear.
- Working from Blueprints: Evolutionary Learning in Design," Sushil J. Louis. Artificial Intelligence in Engineering, Elsevier, 11 335 - 341, 1997.
- 3. ``Seismic Velocity Inversion with Genetic Algorithms," with Qinxue Chen. Proceedings of the Congress on Evolutionary Computation, to appear, July 1999.
- 4. ``A Sequential Similarity Metric for Case Injected Genetic Algorithms applied to TSPs," with Yongmian Zhang. Proceedings of the Genetic and Evolutionary Computation Conference, to appear, July 1999.
- 5. ``Interactive Genetic Algorithms for the Traveling Salesman Problem," with Rilun Tang. Proceedings of the Genetic and Evolutionary Computation Conference, to appear, July 1999.
- 6. "Plasma X-ray Spectra Analysis Using Genetic Algorithms," with R. C. Mancini and I. Golovkin. Proceedings of the Genetic and Evolutionary Computation Conference, to appear, July 1999.
- 7. ``Solving Similar Problems using Genetic Algorithms and Case-Based Memory," with J. Johnson. Proceedings of the Seventh International Conference on Genetic Algorithms, 283 - 290, Morgan Kauffman, 1997.

A more complete list of publications is available from http://gaslab.cs.unr.edu/.

Education and Human Resources - Five graduate students have been supported by this project and completed their Master's in Computer Science. One of them, Judy Johnson, is currently pursuing the Ph.D. at Penn State. We have only recently started offering the Ph.D. in Computer Engineering and two graduate students will be supported in the next academic year under this program. Currently two undergraduate students are also being supported by this project and have been instrumental in implementing the code available from our web site.

Seismological Motivation

This proposal seeks support for software development work that will improve seismologists' ability to use seismogram data from regional earthquake-recording networks. Seismologists have reaped great benefits from the global collection of seismograms over the last ten years by the Incorporated Research Institutions for Seismology (IRIS). To address earthquake source and earth-structure problems at a regional rather than global scale, however, seismologists must gain the ability to access, select, and visualize the seismograms held by the regional seismic networks. A usable set of seismograms from a regional network will typically be larger than a usable set from the IRIS global network, and this project will develop new visualization tools for these larger data sets. Further, regional-network seismogram data stores are distributed worldwide at many different data centers, in many cases maintaining their own data formats and access methods.

This project will also develop, install, and test at several regional networks new seismogram search, data server, and selection methods. Seismologists are now usually able to combine only summary parametric information interpreted from regional-network seismograms, such as wave arrival times. The proposed project will result in the availability of complete seismogram data sets from any participating data center to seismologists around the world. Such availability will be similar to the availability of IRIS global seismograms that seismologists now enjoy.

The software developments for data center servers and user visualization we propose will make heavy use of XML and Java standards. Seismological disciplines are suffering a schism of late due to diverging software developments. Researchers must maintain UNIX platforms to support the available seismogram analysis and visualization packages powerful and flexible enough for their use. Secondary and tertiary students having access to PC computers must use commercialized software limited to basic display and analysis of seismograms. These limits also affect the instructors and institutions of these students. Engineers, architects, and urban planners wishing to use strong-motion seismogram data have an entirely separate set of mostly PC-based software tools.

As a result, no standard for the storage or exchange of seismogram data has been adhered to by many of the regional earthquake or strong-motion networks. Most seismogram data are currently stored in one of about six formats that are quite difficult to mutually interchange. No seismogram-visualization tool is able to read more than three of the formats. The large majority of seismograms are in fact held in the CUSP format (developed in the 1970s by Carl Johnson for use only at seismic observatories), which cannot be read directly by any analysis or visualization tool. The schism between seismologists using UNIX and PC platforms has caused further division of the existing data formats into little- and big-endian varieties.

We propose to simplify access to regional-network seismogram data by registering or extending descriptions for seismogram data types, conforming to standards related to the Extensible Markup Language (XML; Bosak and Bray, 1999). This project will then develop server software to search, extract, and translate the data types held by several of the principal regional networks. We will then implement, install, and test our server packages at these seismicnetwork data centers. The PIs have extensive experience with translating regional-network seismograms, including those in the CUSP format, for their research efforts.

We will develop seismogram search and visualization tools for seismologist users in the Java language. Because Java specifies a standard virtual machine with standard binary data formats across all platforms, the little- or big-endian nature of the user's platform will never demand re-translation of seismogram sample data supplied by any regional network. We will design search tools to allow users with XML-capable browsers to identify and extract sets of seismograms. Server software running as remote methods invoked by users will filter and sub-sample seismograms to supply data sets appropriate to users' capabilities and connection speeds.

Prototype visualization software we have developed for large sets of many seismograms shows that all data-reduction and viewing techniques will be available to all users regardless of their computing platform. Users' visualization capabilities (image sizes and speed of animation, for instance) will scale directly with the size and speed of their computers, with the tools written in Java. The availability of our Java software developments to all computing platforms should bring the seismological researcher, student, and scientist or engineer outside seismology closer to common ground.

Recent Results of Seismic Network Data Access and Visualization

Recent results of the availability of global seismograms from IRIS - The Incorporated Research Institutions for Seismology (IRIS) established a decade ago an archive of global seismograms, with NSF funding. A web page at http://www.iris.edu/proposal/TOC.html points to comments by a number of seismologists on how the availability on the Internet of IRIS global-network seismogram data has revolutionized seismological research. The data are available through the IRIS Data Management Center page at http://www.iris.washington.edu/NEW/HTM/dmc.htm . Peter Shearer demonstrates how global structural studies have benefited by stacking 100,000 seismograms into a summary image tracing the arrival times of heretofore weak and poorly-demonstrated seismic phases. A task impossible before the IRIS data center could supply a global set of seismograms in an identical format from a single data request, Earle and Shearer's (1997) work promises to show the detailed nature and topography of the coremantle boundary and other fundamental earth structures.

The seismological community has been conducting global earth-structure tomography studies for more than 15 years with summary parametric data, usually wave arrival times and earthquake hypocenters and origin times manually interpreted from seismograms. The availability of the IRIS seismogram database has more recently allowed important research that tested structural hypotheses the parametric data leaves undetermined. Revenaugh and Sipkin (1994) could use for example the amplitudes of reflections from the 400-km upper-mantle discontinuity to estimate its physical condition and the presence of melt. Su and Dziewonski (1995) were able to make a detailed characterization of Earth's inner core, and Forte et al. (1995) could begin to estimate variations of the mantle's chemical composition. Niu and Kawakatsu (1997) tested models for mantle convection by examining topography on the mid-mantle discontinuity.

The rapid availability of IRIS seismograms has also allowed accurate studies of the source mechanism of all moderate (magnitude > 5.5) earthquakes. The Harvard CMT Project of Ekstrom and others (represented in the IRIS web page above) has aided local response to severe events such as the in Landers, Northridge, and Kobe earthquakes by accessing a global data set from IRIS, computing mechanism estimates, and identifying the active faults and rupture zones, all within a few hours of the event. H. Kawakatsu (1995) of the Earthquake Research Institute, Japan, and S. Sipkin of the U.S. Geological Survey are also providing rapid earthquake mechanism estimates, but also contribute to a global database of earthquake mechanisms that defines the tectonic regimes of seismically active areas.

Recent results from use of distributed regional-network parametric data - Established prior to the global network of seismographs maintained by IRIS, the more diverse regional seismic networks monitor the earthquake hazard of smaller areas such as single nations, or one to four U.S. States. Arabasz and Malone (1995; with more recent comments referenced on http://www.iris.edu/proposal/TOC.html) explain how the Council of the National Seismic System (CNSS) is a union of 27 U.S. institutions and agencies that operate the majority of

seismographic stations in the U.S., and "has no direct funding but relies entirely on the cooperation and limited resources of its member institutions." Since, as they say, "most networks continue to lack the considerable resources needed to reformat and distribute their data," we propose here NSF support for generalized access to this increasingly valuable body of scientific data.

The regional seismic networks were originally designed for earthquake detection and location by operating mainly high-gain vertical seismographs. To promote thorough event detection, the networks installed up to hundreds of such inexpensive detectors over their regions of interest, yielding excellent geometric coverage that locally far exceeds any of the capabilities of the IRIS or other global networks. A great many seismologists have used arrival time and polarity parameters interpreted from large bodies of seismograms to describe earthquake fault planes and focal mechanisms. Large collections of these summary parameters, from well-sampled regions, allowed detailed earth structure studies below regional networks such as by Hearn (1984) and Humphreys et al. (1984). Studies using summary parametric data continue to break new ground, as in Thurber et al. (1997) describing the presence of deep-crustal fluids within the San Andreas fault from P- and S-wave arrival times.

With the success of the initial studies below individual regional networks, seismologists began to combine summary parametric data interpreted from adjacent networks, to study larger regions than those covered by individual networks. Such combined parametric data sets are not huge, and are generally represented as ASCII text in formats standardized by the CNSS for easy exchange. Hearn et al. (1991) assembled such a data set to reveal crust and upper mantle structure below the eight networks covering the western U.S. Zandt and Ammon (1995), Zandt et al. (1995), and Ruppert et al. (1998) were able to develop compositional as well as structural constraints from similar amalgamated parametric data. Asad et al. (1999) had to combine data from three networks just to investigate the crustal structure of a small area at the intersection of the networks' regions. Comparison of parametric data from even widely separated networks can yield important information from within the mantle, as in Niu and Kawakatsu (1995). The chief difficulty facing such characterization work is the frequent need to gather, view, and re-interpret many of the seismograms for the arrival-time parametric data. Parametric data picked in the course of earthquake cataloging operations at the regional networks may not be adequate for detailed characterization studies.

Recent results from use of seismograms from regional networks - The importance of earthquake-source and ground-shaking characterizations to the hazard-mitigation missions of the regional networks has recently led to increasing installations of high-quality three-component, broadband seismometers by many networks. Local source studies have benefitted, as in Ichinose et al. (1997) for example. Few earthquakes are well enough recorded by any one network, however, and seismograms from global as well as adjacent networks will typically be modeled. Patton and Zandt (1991) modeled seismograms from earthquakes across the western U.S. in an effort at regional tectonic characterization. Ammon et al. (1994) and Anderson et al. (1995) gathered seismograms of deep South American earthquakes from several North American seismic networks to study both event complexity and station site effects in ways very similar to strong-motion studies of local earthquakes. D. Dreger of U.C. Berkeley is providing rapid focalmechanism estimates in response to California earthquakes, by automated modeling of seismograms.

New approaches to earth-structure characterization also require the use of seismogram time sample amplitudes instead of just arrival-time parameters. Surface-wave and receiver-function studies such as by Ozalaybey et al. (1997) require direct analysis of large numbers of seismograms from a large area. Benz et al. (1997) assembled a large number of Lg seismograms to characterize crustal attenuation properties across the U.S. Polarizations revealed by 3-component seismograms yield mineral anisotropy variations, as in Ozalaybey and Savage (1995). Revenaugh (1995) processed regional-network seismograms using summation methods from exploration seismology in an effort to map the scattering character of tectonic features. Chavez-Perez and Louie (1998) used such data and techniques to image blind thrusts below the southern California seismic network.

Expected Results of Regional-Network Seismogram Access and Visualization

Expected results from use of distributed regional-network seismogram data - The few examples mentioned above demonstrate the great value to seismologists of assembling sets of seismograms from more than one regional network. Clear advances in both analysis technique and in understanding earthquake hazards, tectonic processes, and earth structure have been made by those seismologists able to undertake the painstaking amalgamation process at several regional networks. If seismologists did not have to travel to each data center, and did not have to exercise considerable programming talent to translate the often incompatible seismogram formats found among them, they could expend more effort in the interpretation of data.

Advances in the ability of individual seismologists to visualize sets of many seismograms will be crucial to such advances in regional-network seismology. Regional seismograms are experimental data from noisy and poorly controlled environments (the Earth's crust), in which definitive observations may be found in only a very small portion of the recorded data. Seismologists can make new observations in many cases only by rapidly inspecting large quantities of seismogram data, culling out the parts most likely to be definitive. Global seismogram data may have more predictable quality, and seismologists can design data searches that are likely to yield results. Only rapid inspection by visualization, or certain datasummarizing operations such as spectral analysis, reliably yield good results from regionalnetwork seismograms. The current need to design data searches ahead of inspecting data sets also effectively excludes scientists and engineers outside seismology from using global or regional seismogram data sets directly. The proposed project will meet these needs for visualization.

Size and diversity of distributed regional-network data centers - Seismograms from regional earthquake-detection networks reside in CUSP format at the Southern California Earthquake Center (SCEC) Data Center at Caltech (http://www.scecdc.scec.org/ waveforms.html), the Northern California Earthquake Data Center (NCEDC) (http://quake.geo.berkeley.edu/ncedc/ncedc.overview.html) operated by U.C. Berkeley and the U.S. Geological Survey (USGS), the Nevada Seismological Laboratory (NSL) at the University of Nevada, Reno (UNR; http://www.seismo.unr.edu), the University of Utah, the Hawaii Volcano Observatory (HVO) operated by the USGS, the Idaho National Engineering and Environmental Laboratory (INEEL) operated by the U.S. Department of Energy, and the New Zealand National Earthquake Database operated by the Institute of Geological and Nuclear Sciences (IGNS), among others. These data stores amount to about a terabyte in total. Presently, digital seismograms from any of these centers can only be acquired by requesting a copy of a magnetic tape, or by requesting a login account and using UNIX commands to transfer CUSP seismogram files. Organizations such as the NSL, the Univ. of Utah, HVO, and INEEL have not been funded for the distribution of more than summary parametric data such as earthquake location and magnitude catalogs, and so are not yet able to keep seismogram data sets available on line.

None of the regional seismic networks can translate more than a small portion of their seismograms to the SEED format required for archiving by IRIS. Translation to SEED format further requires complete seismograph calibration information, which is not commonly available for regional-network stations. (Uncalibrated seismograms may not be useful for earthquake magnitude studies, but are still quite valuable for source-mechanism and earth structure work.) IRIS, on its part, does not have the resources to increase the size of its data center by an order of magnitude, and archive much of the regional data. This project will create the tools necessary for regional networks to offer their data on-line as they are able, in extensible formats that can include calibrations if available.

Complementary data stores - In addition to the regional earthquake-detection networks, many areas have networks of *strong-motion* instruments that trigger during earthquake shaking to record macroscopic ground motions. Strong-motion seismograms are useful for engineering studies of ground-shaking amplification, and for earthquake source description. Databases of strong-motion seismograms with a query architecture reside at U.C. Santa Barbara (http://smdb.crustal.ucsb.edu/), and at the NCEDC (http://quake.geo.berkeley.edu/db/). Caltech maintains a store of strong-motion seismograms (http://scec.gps.caltech.edu/strongmo/strongmotion.html) for southern California, as does the NSL for a network in Guerrero, Mexico (http://www.seismo.unr.edu/ftp/zeng/GUERRERO/guerrero.html), although these stores must be hand-searched by event date. Even the full database searches result in the transfer of individual seismogram files in one of 4 formats that must be plotted and viewed with additional software.

Another type of seismogram data that will become more accessible after this project are active-source seismic survey recordings (as distinct from passive earthquake-event records). There are large commercial archives of such data for the petroleum industry, and a typical recorded data set may have one or two orders of magnitude more seismograms than the typical regional-network earthquake data set. This project will build on decades of industry experience with the visualization, selection, and summarization processing of such large exploration data sets.

Public stores of active-source survey seismograms are held by the IRIS Data Management Center for the Program for Array Seismic Studies of the Continental Lithosphere (PASSCAL) (http://www.iris.washington.edu/NEW/HTM/dmc.htm), by Cornell University for the Consortium for Continental Reflection Profiling (COCORP; http://www.geo.cornell.edu/geology/cocorp/COCORP_archive.html, by the Australian Geological Survey Organisation (http://www.agso.gov.au/information/structure/isd/database/seis_refl.html; with database queries possible to http://www.agso.gov.au/ngis/locator_info.html), and by the British Institutions Reflection Profiling Syndicate (BIRPS; http://www.ac.uk/bgs/w3/isg/dcat/dcat_subj_list.html#geopprofs).

Both raw and summary seismograms are usually offered in the SEG-Y format established 25 years ago by the Society of Exploration Geophysicists (http://www.seg.org), which is also not

easily converted to the SEED format for IRIS. Much of the on-line survey data are essentially processed summary image sections, where a widely-accepted summarization technique such as common-midpoint (CMP) stacking reduces the dimensionality and size of a data set. This is done by summing seismograms after geometric transformations, akin to stacking radar looks, or to tomographic reconstruction of x-ray or MRI scans. We will apply such summarization techniques to regional earthquake network seismograms, where the data size and a user's connection speed demand them.

Complementary seismogram tools - The tools commonly available to seismologists for plotting and analyzing regional-network seismograms include the free Seismic Analysis Code (SAC) from Lawrence Livermore National Laboratory (http://www-ep.es.llnl.gov/tvp/sac.html), and PITSA (http://www.synapse.ru/software/pitsa/) and JPITSA (a Java version under development) partly sponsored by IRIS. These packages are primarily concerned with accurate and flexible analysis of small collections of seismograms, and not with the visualization or culling of sets of hundreds to thousands of seismograms. SAC and PITSA have to date only been available for UNIX workstations, which forces any seismologist interested in IRIS or other network data to maintain specific computing platforms that may not be easy for students or non-seismologist scientists or engineers to access. It is difficult to translate CUSP-format seismograms to the formats required by these packages. To fully access the large and varied data stores of regional networks, seismologists and others need the visualization and selection tools, and the extensible formats for seismogram data, that this project will develop.

There are several commercial packages for visualizing multi-gigabyte sets of seismograms; the petroleum industry leader is the ProMAX package from Halliburton. A free package known as Seismic UNIX (SU) was written at the Center for Wave Phenomena at the Colorado School of Mines (http://www.cwp.mines.edu/cwpcodes) that matches some of the data-summarization (CMP stacking) processing and visualization capabilities of ProMAX. SU is operated with UNIX shell scripting instead of a GUI. Another shell-based free processing package is our RG system (http://www.seismo.unr.edu/ftp/pub/louie/rg/rg.html). SU, RG, and ProMAX require weeks of training before a seismologist can use them effectively, so they are not practical tools for the classroom or for non-seismologists. We are currently developing JRG, a Java-based application for visualizing and processing sets of hundreds to thousands of seismograms. JRG has a GUI for operational simplicity, and runs on almost any computing platform. (Tested on Solaris, Windows 95, and MacOS; JRG is available through a page explaining a class exercise at http://www.seismo.unr.edu/ftp/pub/louie/class/757/lab1.html.)

Proposed Capabilities

Access and visualization model - We propose to greatly improve the accessibility of regional-network data by developing new seismogram search, server access, and visualization tools. We will rely on new computing technologies and emerging standards to assure that all search and visualization capabilities will be available to seismologists and others, regardless of their computing capabilities. We will create seismogram databases at regional-network data centers, that conform to the new Resource Description Framework (RDF) standard (Bosak and Bray, 1999), with registered Document Type Definitions (DTDs) that will allow any Web browser understanding Extensible Markup Language (XML) to directly search and access data-

center servers. We will implement all software in the Java language, because of its standard virtual machine and uniform binary data formats across all platforms, and its remote-method invocation facilities. Our developments will thus be available for all computers, and we will assure that the data-processing and visualization methods are fully scalable from moderately sized PCs to large supercomputers.

Our proposed seismogram-visualization methods are based on long-established techniques from the petroleum industry. Sets of tens to hundreds of related seismograms are gathered into a "record section." The record is displayed by assigning each seismogram to a row or column of an image, and each time sample's amplitude is represented by a pixel in that row with a color or gray value. Where the pixel resolution of the image is greater than the resolution of the display screen, adjacent pixels are mixed by the visualization software. Where the image resolution is less, the image is overlain by a "wiggle trace" plot of each seismogram. Multi-record data volumes are often represented by animating the images of individual records.

Distributed database search by seismologists and engineers - We will develop Resource Description Frameworks (RDFs), as described by the World Wide Web Consortium (W3C; at http://www.w3.org/TR/PR-rdf-syntax/), for seismogram and summary parametric data held by regional seismic networks. The RDFs will allow each data center to catalog its data stores, and publish in a standard way their data formats and file access schemes. After developing the seismic RDFs during the first project year, we will register them with W3C, RDF discussion lists like rdf-dev (http://www.mailbase.ac.uk/lists/rdf-dev/), and organizations such as Schema.net (http://www.schema.net) for comment by other database developers. We will present our RDF data-server frameworks to other seismologists operating regional networks at scientific conferences.

Our seismogram RDFs will bear some similarities to audio multimedia and geographic information system (GIS) RDFs listed at Schema.net. Some of the seismogram search capabilities we will develop are essentially geographic, and we will adopt where possible the geographic RDF cataloging schema being implemented by the GILS project at the USGS (http://www.gils.net/), by the ESIP Federation for the Earth Observation System sponsored by NASA for remote-sensing images (http://harp.gsfc.nasa.gov/~eric/), and by the California Environmental Resources Evaluation System (http://www.ceres.ca.gov/) for GIS data. We will develop during the first and second project years a set of XML DTDs that describe earthquake and active-source survey seismogram searches and search results, extending and generalizing in a way the search facilities at the IRIS Data Centers (http://www.iris.washington.edu/NEW/HTM/dmc.htm). Users' XML-capable web browsers will use the search DTDs to query the RDF databases at data centers.

During the first project year we will develop data-store scanning and RDF cataloging software, and query-response methods, in Java for the regional-network data centers. These results we will install at 3-5 cooperating regional networks during the second and third years. RDF architecture will be based on geography and earthquake size and time. Unlike an IRIS Data Center search, a user search with our methods will not return pre-made GIF images of seismogram plots. The user's XML browser will instead invoke methods from our JRG seismogram-visualization package that will offer to download records to the user's computer after varying degrees of summarization by the server, keyed to the likely transfer time given the user's connection speed. These browser and database enhancements to the JRG package we will

code in Java through all three project years. We will also present our search and visualization packages for user evaluation at seismological conferences.

As users search our RDF databases at the installed data centers, the servers will archive the activity of all users (anonymously). Data users will also be encouraged to fill out short questionaires evaluating our access system, and describing their success with finding and using regional-network seismograms. During the second project year we will build the questionaires and the search and result database routines. In the third year, we will develop neural-network methods to train servers to make effective responses to classes of typical users. For example, the seismologist who identifies an interest in source-mechanism modeling can be offered a set of seismograms selected for good azimuthal coverage. A secondary-school student stating interest in the latest large earthquakes can be offered a few high-quality seismograms together with parameters for textbook-like displays.

Data center security - We will build into the server software from the beginning security methods that data centers can invoke to allow access only to registered users. The RDF databases will include sensitivity and security tags that can allow search routines to miss or deny access to certain data sets. The user-activity databases and neural-network-trained search engines will provide further opportunities for data centers to allow or deny access very specifically. As we develop our data-server systems, the possible security procedures will also develop.

Conversion of multiple formats - During the first project year, we will enhance our JRG seismogram-visualization package with new methods to directly read and convert the high-priority seismogram-storage formats: CUSP; SEED; SAC; SEG-Y; and CSMIP. We will write these conversion routines as remotely-invocable Java methods for the servers. During the second and third years we will create routines for server conversion of other popular seismogram formats: CSS; AH; RefTek, SEG-2; SEG-D; SSA; and USGS-SMC. Not all of these routines will be in portable Java code; some may convert to a SEED or SEG-Y intermediate format.

Users will remotely invoke server methods to create serialized Java objects for download from one of the existing seismogram formats. Data centers will thus be able to easily incorporate new data formats, simply by providing a SEED or SEG-Y conversion routine for their own machines. Prototype C code is in our RG package for reading CUSP and SEG-Y, and our JRG package has prototype Java code for SEG-Y, and the serializable class for data transmission.

With seismogram data-transmission classes described and extended by XML DTDs, new tags from new data formats can be added without affecting the ability of the seismogram-visualization package to display the data. The serialized binary data will contain the minimum information needed to describe the seismograms; information about source and receiver geometry, earthquake source parameters, station calibrations, previous processing, time picks, etc. will be carried by the extensible XML tags and described by the DTDs. An advantage of having this information in XML tags will be that difficult-to-read properties such as those in a binary SAC or SEG-Y trace header will be plain readable text in XML.

The ability of our software to deliver disparate and distributed data sets to remote users will be similar to the capabilities of UCAR/Unidata's Distributed Oceanographic Data System (DODS; http://www.unidata.ucar.edu/packages/dods/index.html). Unlike DODS, we will re-code existing data-translation routines from C or Fortran into Java, and rely on remote method invocation (RMI) instead of the http protocol to deliver data to the user. Our servers may be more

difficult to install than a DODS server, but ours will not be restricted to particular platforms.

Data server record gathering, selection, and sorting - Servers at regional-network data centers can provide sets of hundreds of seismograms to remote users only if the users can invoke data-selection and summarization processes, on the server machines, to reduce the amount of data transferred. Optional reduction of data volume on the servers will provide the scalability of our visualization developments from small to large computers. Lossless compression of binary seismograms usually do not reduce data amounts by even half. Lossy compression schemes used in audio technologies are not acceptable since seismograms are usually analyzed further after transfer, and seismologists rely heavily on having accurate, uncorrupted data. The amounts of data transferred in response to typical user requests can be drastically reduced through the use of data-selection and summarization processes that are well understood and accepted by seismologists.

The SEED, CUSP, and SEG-Y files archived at regional networks typically contain complete data sets recorded in response to earthquake or triggered events. Such files will include auxiliary timing and calibration channels, three-component data wherever available, and all stations in the regional network even if inoperable at the time. Our seismogram-search procedures will allow the user to select just stations within a desired distance range, for instance. Earthquake-source modelers might only want to transfer data from 3-component instruments, and crustal-structure analysts might only want the vertical components. Timing and calibration channels would be employed in the process of data conversion on the server if necessary, and would not be transferred.

Time windowing will also reduce the amount of data transferred to users. Most archived files encompass some large time period to save any data of interest to any user. The query and search methods will allow users to select the time windows they need for their analysis; including windows in time reduced by epicentral distance if useful. The server routines will also be able to sort the seismogram records to arrange them by distance, source depth, source or receiver location, and other geometric parameters. We will include the channel and distance selection criteria and time-windowing in our server software during the first project year; sorting procedures we will add in the second year.

Data server filtering and stacking summarization processes - Another simple way to greatly reduce data-transfer sizes for many users is to lower the frequency content of the seismograms. In the first year we will include low-pass filtering and time-sample decimation routines in the server software. Many CUSP data stores are sampled at a 100 Hz frequency; many users will resample to 10 Hz or below, since they may not be interested in frequencies above 5 Hz.

Many seismologists will also add many seismograms, sample by sample after shifting by a source or wave arrival time, to stack the effects of many events at one station, or to stack many recordings of one event. Sometimes the stacked seismograms are selected on a geometric criterion, grouping nearby stations or earthquakes, or "mixing" traces having nearly the same source-receiver distance. Reflection survey data are stacked into midpoint areas after a geometric transformation known as NMO correction. The NMO-correction and CMP-stack process reduces the dimensionality and hugely reduces the size of a data set, although such analyses depend on additional information. With a regional seismic-velocity model, a large data set can be stacked with Kirchhoff summation into a reflectivity section, as in Chavez-Perez and Louie (1998).

During the second project year we will include simple stacking in our server routines; more sophisticated summarizing analyses we will include in the third year.

Record display and animation by seismologists and engineers - Seismogram records that have been selected, sorted, time-windowed, filtered, and possibly stacked will not be too large to transfer even to users with modem connections. Users with faster connections will be able to transfer a larger and more complete data set from a server, and apply any summarizations themselves. The summarization methods will be built into Java classes, and so can be executed remotely on servers or locally on the user's machine. Our data-search procedures will also suggest to users the summarizations needed to allow effective visualizations of the record, given the capabilities of the user's machine. During the third project year our neural-network training of server search tools by user access databases will be configured to produce hints on data summarization for users.

Louie (1990) compared wiggle-trace, grayscale-image, and color-image visualizations of seismogram records. Our current prototype JRG visualization package now allows automatic pixel mixing of gray and color images where the image resolution is higher than the display resolution. During the second project year we will add optional wiggle-trace representations, and in the third year we will allow wiggle-trace representations to appear automatically when the user scales the image to a lower resolution than the display.

Louie (1990) also compared animated to volume views of multi-record seismic data sets. For the purposes here of selecting good-quality seismograms from regional-network data stores, and for interpreting arrival times, animations are simpler to work with than volume-rendered 3-d views. The JRG prototype now provides smooth multi-record animation of data volumes that will fit into the user's RAM, not more than 20 megabytes on most machines. During the third project year we will rework the animation system to use a memory frame cache, rolling through a data volume saved on hard disk. This will allow animation and processing of much larger data volumes. In the third year we will also augment the visualization system with seismic-phase identification and time-picking capabilities. The user's interpretations and data processing sequences can be recorded in XML documents (backed by an appropriate DTD) and sent to others, or even back to a server for inclusion in the usage databases.

Examples of prototype seismogram visualization system - Our prototype JRG software system for seismogram visualization and animation has a simple GUI, and includes limited data-translation and summarization-processing facilities. Written in Java for portability, JRG is available from a course exercise at http://www.seismo.unr.edu/ftp/pub/louie/class/757/lab1.html, or from the java archive at http://www.seismo.unr.edu/ftp/pub/louie/class/757/java/jrg.jar . We have several illustrations of JRG's capabilities that are not included in this proposal for space reasons. The illustrations, between 48 and 80 kilobytes in size, are on-line at the locations given.

The first illustration shows a visualization of a sequence of 249 earthquake seismograms, recorded at a single station, running on Sun Solaris. The seismograms are plotted with the time axis pointing down, and their amplitude values represented as colors in the image (red is positive amplitude, blue is negative, and white is zero amplitude). P- and S-wave arrivals are evident between 4 and 8 seconds, although the seismograms have not been sorted by any geometric property such as epicentral distance, and are arranged in order of event occurrence. The trace-equalization gain evens out the rms amplitude of each seismogram. The superimposed dialog controls the visualization. The seismograms were acquired in CUSP format from the New

Zealand National Seismograph Network's database, courtesy of T. Webb and S. Bannister of the IGNS. (http://www.seismo.unr.edu/ftp/pub/louie/proposals/vis/record.gif)

The second illustration shows a spectral analysis of a similar sequence of seismograms, running on Windows 95. A trace-mixing data-summarization process is also employed. Variations in the overall energy of seismic waves received at the station are apparent as the sequence progresses. The analysis was performed by selecting processes from the Methods menu on the receiver gather and spectrum image windows. (http://www.seismo.unr.edu/ftp/pub/louie/proposals/vis/spect.gif)

The third illustration shows an animated visualization of the results of a software process, running on a Macintosh. The process computes 12 megabytes of synthetic wave travel times to points in a crustal section, used in a prestack migration data-summarization process as in Chavez-Perez and Louie (1998). Animating the time-data volume allows rapid consistency checks, and identification of areas where the migration will image reflected refractions instead of wide-angle reflections. (http://www.seismo.unr.edu/ftp/pub/louie/proposals/vis/animate.gif)

Testing on new NSL data center - Our initial installation and tests of server software and RDF seismogram catalogs will be at the Nevada Seismological Laboratory (NSL). During the first project year we will acquire and install the proposed \$20,000 of server hardware. This acquisition of a 144-gigabyte server system will allow substantial on-line archiving of Nevada data for the first time. We will begin to implement an RDF-format catalog and databases for the USGS-supported Western Great Basin Seismic Network (WGBSN) CUSP archives, the 1990-1998 Nevada State Digital Network, 36 new broadband stations supported by the U.S. DOE and the W. M. Keck Foundation, and an NSF-supported strong-motion accelerograph network in Guerrero, Mexico.

The NSL is implementing a new realtime recording system based on Earthworm (a USGS system) from Kinemetrics/BRTT, called Antelope. This UNIX realtime system will have its own database server (not supported under this project), isolated from the proposed RDF server for reasons of data-recording reliability, since the RDF server may be bogged down by users' data-summarization processes. During the second project year we will begin automatic updates of the RDF catalogs and databases from the Antelope system.

Utility to students in geological sciences and engineering - Access to our seismogram server and visualization developments will only require freely-available XML browsers and Java runtime environments. Therefore we anticipate that secondary, tertiary, and graduate students (and their instructors) will be able to make use of regional-network seismograms in a great variety of learning environments. Our server-access databases and usage questionaires will address educational as well as scientific uses of seismograms. During the third project year we will enable the servers to provide guided access and display of regional-network seismograms to students and instructors who have identified themselves as such. The guided access will automatically provide data expected to be suitable for educational purposes at each level.

Distribution of software products - During all three project years we will fully document our software developments with on-line Web tutorials and reference manuals. Both server and user software, including source code, will be available for public download (from http://www.seismo.unr.edu) throughout the project period. During demonstrations at scientific meetings we will hand out diskettes and CD-ROMs with manuals and software to both users and regional seismic-network operators. We will work closely with other regional networks around the world to assist them in using our developments and creating their own seismogram databases with RDF catalogs.

Work Plan

Task	Year	Task	Year
Develop seismic RDFs	1	Register RDFs for comment	1,2,3
Present RDF schema to seismologists	1,2,3	Develop seismogram search DTDs	1,2
Develop RDF server cataloging software	1	Develop server query-response software	1,2
Install server hardware at NSL	1	Install server software at NSL	1
Catalog NSL seismogram stores	1,2	Develop auto-update of NSL RDFs	2,3
Install server software at 3-5 networks	2,3	Install user-software devel. platforms	1
Develop JRG browser and search links	1,2,3	Present JRG user package to seismologists	1,2,3
Develop user questionaires	2	Develop server search & result databases	2
Develop neural-network guided searches	3	Develop JRG import of high-priority formats	1
Develop conversions for other formats	2,3	Develop and register seismogram DTDs	1,2
Implement channel selection on servers	1	Implement time windowing on servers	1
Implement trace sorting on servers	2	Implement filters & resampling on servers	1
Implement stacking on servers	2	Implement server summarization analyses	3
Develop NN-guided data summarization	3	Implement wiggle-trace visualization	2
Implement wiggle-trace automation	3	Implement JRG animation cache	3
Implement picking in JRG	3	Develop on-line tutorials and manuals	1,2,3
Demonstrate and distribute software	1,2,3	Advise data centers on RDF development	2,3

PI Experience With Visualization of Distributed Data

Research experience with visualization - J. Louie has been investigating visualization techniques for seismograms since 1987. At a Society of Exploration Geophysicists International Meeting Workshop on Visualization he compared methods for rendering and interpreting 2-d seismic records and 3-d seismic data volumes (Louie, 1990). An analysis of reflection-survey data across the Garlock fault in So. California (Louie and Qin, 1991) relied heavily on both animated and voxel views of 3-d prestack field records and synthetics to trace an earthquake fault to the surface. Chavez-Perez et al. (1997; http://www.seismo.unr.edu/htdocs/students/CHAVEZ/paper1/paper1.html) used similar visualizations with Death Valley, Calif. reflection data to image for the first time in true depth a listric normal fault. Chavez-Perez and Louie (1998; http://www.seismo.unr.edu/htdocs/students/CHAVEZ/paper2/paper2.html) developed methods to convert and summarize regional-network earthquake seismograms in CUSP format from So. California, and imaged blind-thrust fault structure below the Northridge earthquake. Louie et al. (1997) and Asad et al. (1999) used velocity-modeling processes on amalgamated summary parametric and seismogram data from four regional networks to reveal crustal structure.

S. Louis has developed a Java visualization framework that makes it easy to create views to visualize new kinds of data. His experience in Java and seismic applications is shown with several conference publications on "Seismic Velocity Inversion with Genetic Algorithms," found at http://www.cs.unr.edu/~sushil/ papers/conference/conf.html . He has investigated the application of genetic algorithms to seismic inversion, finding there is essentially no difference in quality of results when comparing GAs against simulated annealings (SAs). However, GAs are easily parallelizable, so there they have an advantage.

Teaching experience with visualization - J. Louie has taught for 5 years a Scientific Visualization Seminar and Workshop course to graduates, attracting students from many departments at UNR (http://www.seismo.unr.edu/ftp/vis/intro.html). His Seismic Imaging graduate course now employs Java-based seismogram modeling, analysis, and visualization tools (http://www.seismo.unr.edu/ftp/pub/louie/class/757-syll.html). Examples of voxel visualization of 3-d seismogram volumes he includes in several courses (http://www.seismo.unr.edu/ftp/pub/louie/class/757/cvmod/cvmod.html; http://www.seismo.unr.edu/ftp/pub/louie/class/hydro/nmo.html), and also examples of animation of such volumes (http://www.seismo.unr.edu/ftp/pub/louie/class/hydro/seis-princ.html).

Experience creating Internet access to distributed parametric data - J. Louie oversaw the development of the NSL's Web site five years ago, and trained the undergraduate students who implemented it. The site provides

unique access to automatically-updated maps of recent events (http://www.seismo.unr.edu/Catalog/fing.html), the Nevada Broadcast of Earthquakes (http://www.seismo.unr.edu/Catalog/nbe.html), and a geographic and chronological event-search and mapping facility (http://www.seismo.unr.edu/Catalog/catalog-search.html). G. Ichinose of the NSL developed the JavaWorm interactive event-mapping facility (http://www.seismo.unr.edu/fttp/pub/ichinose/JavaWorm) for Java-enabled browsers.

Experience creating scalable seismogram analysis and visualization - J. Louie began creating the RG seismic-processing package at Caltech in 1982, and it is now available as a special-purpose alternative to SU and other free processing packages for UNIX machines (http://www.seismo.unr.edu/ftp/pub/louie/rg/rg.html). He recently began development of the JRG seismic-processing package in the Java language. It was originally created to meet the seismogram-analysis programming, animation, and visualization requirements of his Seismic Imaging course (http://www.seismo.unr.edu/ftp/pub/louie/class/757/lab1.html). By adding a graphical user interface to allow students to read and process data collected by refraction seismographs, JRG's application has been extended to senior-undergraduate-level courses. A detailed tutorial on using JRG for one type of seismogram analysis is at http://www.seismo.unr.edu/ftp/pub/louie/class/vspect/vspect.html.

Letter of Interest from the New Zealand National Seismograph Network

John N. Louie, Assoc. Prof. of Seismology

Nevada Seismological Laboratory, University of Nevada 174, Reno, NV 89557-0141 USA 9 June 1999

Dear John,

I am pleased to write in support of your proposal "Access and visualisation of large distributed regionalnetwork seismogram data sets" to the U.S. National Science Foundation's new programme on large scientific and software data set visualisation. Your objective of enabling co-operating seismologists to access the high-quality seismogram data stored at regional-network data centres world-wide will assist us in disseminating seismological data to engineers and scientists, and also those outside the seismological community. Your proposal to develop seismogram search and visualisation tools will also assist us in our research into earthquake hazards and hazard mitigation through providing easier access to our own data and more sophisticated tools for its analysis.

If your project is funded, we will invite you to install and test your server software on our database of digital seismograms held here in the CUSP format. We will need to place two restrictions on access to data: 1) data collected as part of a specific project will be embargoed for a time so that the people who collected the data can work on it; 2) people wanting to work on local New Zealand problems will need to work with our staff. People working on global problems can have free access to all data that is not embargoed – some data has already been used in this way. We would also like you to measure the impact of your data server's operations on our computing environment, and will need to moderate access if it is causing response problems. Our staff will keep you informed about our own data server developments, and we are looking forward to benefiting from your development of scalable and portable software.

New Zealand's national seismograph network is in many ways comparable to the regional seismic networks in the U.S. run by organisations such as the Nevada Seismological Laboratory. The main differences are it that it is more sparse and very little of it has real-time telemetry. On the positive side, many of the standalone stations are 3-component, and the data are gain ranged so are of relatively high quality for short-period data. With more than 30 permanent stations, we record about 14,000 earthquakes each year. We also conduct temporary seismograph deployments, often with international participation, and data from those experiments are usually merged with National Network data and processed through the CUSP system to give one complete, integrated database. The database currently contains over one million digital seismograms in the CUSP format, amounting to some 30 gigabytes of data that we keep on-line.

Our Institute is participating in GARNET, a data sharing initiative for the western Pacific region. We will also be contributing to the new international nuclear test detection network based in Vienna, once two of our stations have been upgraded to their auxiliary station standard. These kinds of initiative create a workload in fulfilling data requests, we thus hope your work will be able to extend our ability to co-operate in such efforts at regional scales, and with scientists and engineers from other disciplines.

Terry Webb Programme Leader, Geological Hazards Institute of Geological and Nuclear Sciences, Lower Hutt, New Zealand