COSMOS Strong-Motion Virtual Data Center

MELINDA SQUIBB
RALPH ARCHULETA
JAMISON STEIDL
University of California, Santa Barbara

ABSTRACT

The COSMOS Strong-Motion Virtual Data Center (VDC) is an unrestricted web-based search engine for access to worldwide earthquake strong-motion ASCII data files (http://db.cosmos-eq.org). It provides an interactive resource for research and practicing earthquake engineers, earth scientists, and government and emergency response professionals.

Users have a wide range of access options; using the VDC, they may search for specific characteristics of the data, view data in a geographical perspective, preview records, compare response spectra with design spectra and download the data and metadata of most interest to them.

The VDC continues to expand and significantly improve the accessibility and the use of strong-motion records worldwide. A number of initiatives are pending to streamline data acquisition and delivery, fill out incomplete metadata and generate missing data products, standardize and simplify data formats, modernize and enhance the web interface, and provide access to tools for file creation, preview and processing.

CURRENT FACILITIES AND PROCESSES

The VDC was originally developed and maintained under grants from the Southern California Earthquake Center and the National Science Foundation, and is currently funded, under COSMOS sponsorship, by the USGS and the CGS. A Working Group composed of COSMOS members and including representatives from the USGS, the CGS, the U.S. Army Corps of Engineers, the U.S. Bureau of Reclamation, and the professional engineering community sets the development agenda and monitors the progress of the VDC.

Data

As of May 1, 2006, the VDC database contains the metadata for 522 earthquakes, 3150 stations and 26,876 time histories (including the URL addresses for each strong-motion acceleration, velocity, displacement, Fourier and response spectra file) from the U.S. and 14 other countries. Although the VDC provides access to a considerable amount of earthquake data, as a metadata database, the VDC database is small. Thus database disk space and retrieval latencies have not yet been a pressing issue for the VDC. The VDC has used Oracle and MSSQL, and is currently running in Postgres, all of which have been adequate for the task.

Data files are ASCII strong-motion files, and as such are relatively small in size and small in number. The VDC currently distributes data from earthquakes of magnitude 5 or larger in areas of high seismicity and 4.5 in areas of low seismicity. In addition, since the VDC is particularly directed at the needs of those examining the effects on structures, files representing events at depths greater than 100km have usually been excluded. The VDC includes an ftp server for distribution of certain legacy files and files from providers who have limited computing resources, but the vast majority of data files reside on the servers of the original data providers, so that the disk space for file storage has also not been an issue. The latencies for redirecting data from the data providers have been acceptably small, but server availability or lack of notification of changes in data providers' servers has occasionally been an issue. Currently this problem is managed to some degree with scripts that test for availability of data on a weekly basis.

Although there is every effort to distribute data files as soon as possible, they are not distributed in real time. Most users of the VDC prefer processed data (although raw data is also provided), and most data providers prefer to do their own processing, which delays distribution somewhat. The Strong-motion VDC model requires merely that the data providers place their data on their own ftp server; the data providers are not required to modify their data for uniformity or actively transmit it when available. This passive model has encouraged many providers, especially those with limited resources, to contribute their data to the VDC. Nonetheless, the lack of a streamlined push or pull process for newly available files, the abundance of data formats and lack of standardization even within formats for some providers have hampered attempts at automating the process of metadata ingestion into the database. The current process is not a scaleable one; there are now more than 25 data formats represented in the VDC and that number continues to increase. Although the VDC, in cooperation with the COSMOS Strong-motion Program Board, is developing a format that will accommodate both U.S. and international data, dozens of data providers across the world cannot realistically be expected to instantly change their formats and modify their own software to suit the VDC, or to go back and convert legacy data. Conversion to common formats will only be adopted with gentle pressure from data distributors and users; and when adequate, free and easy-to-use tools for conversion and processing are available.

An additional drag on turnaround time is the creation of files of acceleration and response spectra used to preview data at the website. Currently, acceleration plots are created for all data, and response spectra files for all provider-processed spectra, corresponding to approximately half of all of the records currently available. In the future, the VDC plans to add plots for velocity and displacement, and make all plots rescalable. The two types of files currently available are created by scripts, but the process would be facilitated by a standardization of formats.

Search Mechanisms

There are several methods for searching the metadata parameters: a map interface, earthquake and station lists, a basic parameter search page and an advanced parameter search page.

The Map Interface displays earthquakes and stations using either a simple two-dimensional outline map or, for California data, an outline map with fault lines, each dynamically drawn in real-time. Users may reconfigure the map by entering latitude and longitude ranges, zooming in, clicking on a station or earthquake symbol to transfer to station or earthquake pages, respectively, or highlighting the stations reporting a selected earthquake. Recently a link to a GoogleEarth KML file was added to the website. Users must

download GoogleEarth to their own machines and drag the KML file to the application. The KML file updates its VDC data every 24 hours or on demand. Once the file is loaded, GoogleEarth displays VDC stations and earthquakes and each icon links to the corresponding station or earthquake page on the website, but the application does not currently interact dynamically with the website. While the results of any search may be drawn dynamically on the VDC map in a bounding box corresponding to that data, the GoogleEarth map only displays all stations and earthquakes and the user must narrow the geographical area of interest on his own.

The Earthquakes Page lists earthquake name, magnitude, number of stations and data provider for all earthquakes available through the VDC by region, with a drop-down list of regions at the top of the page for quicker navigation. The Stations Page lists owner and station name for all stations available through the VDC by region, also with a drop-down list of regions at the top of the page.

The Basic Search Page allows the user to enter the most common parameters. The user may also tailor the output to reflect station information only, earthquake information only, or all data, for the result set. The Advanced Search Page allows the user to query and recover almost every field in the database. The user may select an html table or a station page as the output of the advanced search, or download the metadata as an rtf file.

Downloading Data

Data files may be downloaded in a batch (as tarred and zipped files) or individually. The USGS and the two Japanese networks, K-net and Kik-net currently provide zipped archives of their data files for each earthquake. The link to download these is available on the Event Page for all events for which these are available. The VDC will also be providing zip archives of all files on its own ftp server in the near future, and will encourage its other data providers to provide zip archives for each earthquake of their files as well.

The VDC also provides downloads of files on an individual basis or as an archive of files resulting from a search via a shopping-cart mechanism. Each type of search generates a page of all stations meeting the search criteria, and includes data about each record, e.g. pga, pgv, orientation for each sensor component, site geology, links to preview plots and a map and checkboxes to download each component, or each station's data or the whole page's data. After the user has selected the data, he enters his email address and is directed to a download page. This page has all files pre-selected and the user may click one button to download all files as a succession of zip files (30 files per zipped file), or may view and download any file individually.

The files are free and unrestricted, although the VDC requests that users cite the data provider and the VDC in any publications. One disadvantage of the virtual nature of the VDC for the user has been the plethora of file formats. Users must find or write their own software tools to process and view data once downloaded and they may have to contend with many formats for a given project.

In addition to downloading strong-motion data files, users may select metadata fields in the Advanced Search page and display them as a table of values or download them as an rtf file. The later is importable into Excel.

PENDING IMPROVEMENTS AT THE VDC

Website

The VDC web pages are currently generated using perl CGI scripts. Perl is free, widely available for all platforms, with drivers for most databases and many modules for doing the operations at the VDC: downloading and zipping data, creating graphics, etc. However, perl is slow compared to current implementations of Java in a servlet container like Tomcat or an application server like JBoss. In addition, VDC scripts were originally written without regard to ease of maintenance and are often cumbersome to update. The VDC is currently testing a completely redesigned site that will use Model-View-Controller design, which separates more cleanly the user interface component from the database access, program logic and site navigation components. This design model is used widely to facilitate periodic website maintenance and future modifications and results in a more robust and scaleable website. In addition, the site will use CSS stylesheets for a more consistent appearance and follow standards for ADA compliance. The migration to java will also facilitate incorporation of technologies less well supported in perl, including web services, xml processing and mapping middleware.

The redesigned station and event pages will have more information on each page, reducing the need for searches. The VDC is also planning to add parameters to aid in searches of structural data. Response spectra preview plots at the VDC are currently configurable and allow the overlay of various design spectra, but must be reentered at each visit to the site. The VDC plans to allow the user to store sets of his design parameters for future use. The input from members of the engineering community for these web pages in particular has been invaluable.

The VDC continues to monitor and evaluate the performance of the site with respect to current conditions and recent technology, and assess needed upgrades or changes to hardware or software, including the server, OS and database systems.

Additional Data: Strong-Motion Files, Response Spectra, and Interoperability

The VDC is in the process of incorporating the remaining legacy data from the NOAA/NGDC data disks and has recently received permission to incorporate data from New Zealand's Institute of Geological and Nuclear Sciences from 2000 to the present and to harvest data from their website on a continuing basis. We previously had incorporated New Zealand data from 1966 to 1999. New partners for data are being actively recruited.

In addition, the K-net and Kiknet networks in Japan have offered to process their data and add response spectra files to their server in the near future, which the VDC will add to its website as soon as they become available. With the addition of these data and the NOAA data, the vast majority of records available through the VDC will have response spectra. The COSMOS VDC Working Group is overseeing the processing of the remaining records to produce preview spectra on the website, using guidelines developed by COSMOS.

In conjunction with the COSMOS Geotechnical VDC, the VDC is developing a map interface that will allow users to access the GVDC borehole data in the vicinity of Strong-motion stations. Jennifer Swift's paper for this workshop will address this in detail. The VDC is also working to establish a closer relationship

with ANSS, CSMIP and USGS to streamline access to data, to add legacy data with a magnitude less than 5.0 and to provide access for VDC users to related data and tools. Interoperability with other data sites is an efficient use of resources, and benefits the user population with data they might not otherwise have been aware of or which might be difficult for them to correlate with their strong-motion data from the VDC. With increased interoperability, users will only have to search once for geographically collocated data.

Common Data Formats and Tools

The VDC, in conjunction with the USGS National Strong Motion Program, the CGS Strong Motion Instrumentation Program and practicing structural engineers, is developing two alternative formats for strong-motion data, one based on name-value pairs and the other XML. The formats are in addition to the traditional COSMOSv1.2 text format developed in 2001. The new formats will be used both internally, to streamline the data ingestion and format conversion process; and externally, to provide a flexible format that is more amenable to international data and can more easily adapt as new equipment is introduced, new metadata parameters arise and new technologies evolve.

Both formats are designed to be human readable, so that users will not have to refer to a format document to understand the content, although there will be defining documents. Both formats specify more metadata parameters than previous strong-motion formats, and explicitly state which parameters are required, conditionally required or optional. It is hoped that by including all desired metadata parameters, data providers will be encouraged to supply more of that metadata with their files.

Both formats will include text parameter values rather than using codes to refer to external look-up tables, since those tables may be difficult to locate in the future. In previous formats that used codes for such things as sensor types or networks, the tables too often became obsolete as soon as they were published and were cumbersome to update. These new formats will have an 'other' category for every such value, and a simple process for updating the tables contained in the defining documents. Thus, rather than relegate custom or new equipment to a comment field, as in previous formats, the data values will be associated with the proper tag, so that they are computer parseable and can be processed for databases and file conversions. In addition, there will be a number of comment fields, associated with particular components of the header (e.g. event, sensor, processing), to add the metadata that cannot be simply described in a short text field. Older formats usually have a single comment field that might contain information regarding any aspect of the file, and thus could not be easily incorporated into database systems.

Both formats will contain more data crediting agencies responsible for various components of the file. It is often true of strong-motion data that earthquake parameters are calculated by one or more agencies, which may be independent of the station owner(s), the data processors, and the agency that assembles the file itself. Each of these can be clearly identified in the new formats.

Tagged Format

The first format, a tagged format, is an ASCII format file whose header consists of tag-value pairs, one per line, followed by data, also one value per line. The total file size is approximately 10% larger than the original v1.2 format, but the importance of file size per se has diminished considerably in the last decade. After the first line, which defines the format name and version, tags may be in any order, and all optional and many conditional tags may be omitted entirely, thus allowing providers to tailor the headers to their preferences.

To import the data into Matlab or Excel, the user need not write any software, just delete the header and the trailer line and import the resulting file directly into either application. The VDC, however, will be supplying software tools to export the files into those applications while retaining the appropriate metadata for generating plots and processing the data.

The VDC will create and make available a converter for each format, including legacy formats, into this tagged format. A java program that converts from COSMOS v1.2, SMIP's format and USGS's SMC format is already in beta, and awaits final modifications to and acceptance of the tagged format. The VDC will also make available to the general public a GUI application that will convert from this tagged format to various standard formats in use today. The VDC is aware that users may possess software that requires a particular legacy format, and we can assist users by providing tools to convert any files they download into their required format. Necessarily, however, because the tagged format contains more metadata parameters than traditional formats, the file converted to the legacy format must deposit the remaining parameters into a comment field.

XML Format

The second format is an XML version of the tagged format. Although this format is, strictly speaking, human readable, it is less so than the tagged format, and is envisioned largely as a format for exchange of data between computer applications. XML is today the standard for data exchange, as evidenced by its prominence in ISO standards, web services, and configuration files.

XML files are not only computer parseable, but validation of the data values in them is already built into the format via the defining schema. Moreover, once a file is in XML, it is more easily converted to another XML format. This may be useful, for example, for ArcGIS, which can use GML files as input, and GoogleEarth maps, which uses KML files, since both GML and KML are XML format standards.

The VDC anticipates creating a conversion script for each provider format into this XML, then converting each strong-motion file into XML for internal use, so that only one set of processing scripts need be written for ingesting data and preparing plots, rather than the multiple sets of scripts used now. These scripts would also be publicly available for those who prefer to use this format.

The VDC will also give data providers an application to create tagged files from scratch or modifying ones that already exist. This will facilitate data importation from databases and from files containing individual components for a strong-motion file. For example, a data provider might use the GUI front end of this editing tool to create the metadata for an event, then drag and drop the event and the corresponding station data into each successive file they create through the editor; or they might call the command-line version to script the building of files. The editor, like the converter, would also contain a module for viewing the data. Because the XML schema already contains the apparatus for validation, the editing tool will utilize the schema itself to structure the application. Because the schema will already contain the tables used in the format, these may be displayed as drop-down lists in the GUI application. Such an application may also have an auto-updating function, to add values to tables, for example. Providing tools to convert, view and process files in this tagged format is key to adoption on a wider scale.

CONCLUSION

The VDC is a robust portal for strong-motion data that has been in continuous operation since the mid 1990's, and has become a widely recognized and used resource for engineers, scientists and students. Periodic assessments of new technology, utility, speed and cost have occasioned modifications and additions to the VDC. Thus, the site has changed platforms and underlying databases, added response spectra and other data products, simpler file downloading, and configurable design spectra. Recent evaluations are addressing problems of the scalability of its processes through a common format; enhancements to its user interface, especially its map interface, in light of recent developments in mapping technology; and interoperability with related sites through agreements with the GVDC and other partners in the fields of engineering and seismology.