OPTIMIZING DATA ACCESS AND AVAILABILITY FOR SEISMIC CALIBRATION RESEARCH

Michael L. Begnaud, Richard J. Stead, Julio Aguilar-Chang, and Hans E. Hartse

Los Alamos National Laboratory

Sponsored by National Nuclear Security Administration
Office of Nonproliferation Research and Engineering
Office of Defense Nuclear Nonproliferation

Contract No. W-7405-ENG-36

ABSTRACT

The Ground-based Nuclear Explosion Monitoring Research & Engineering (GNEM R&E) program has made recent advances in optimizing data access and availability for seismic calibration research. Some of the most challenging tasks of maintaining functional and accessible data warehouses are the development of software to automate the continuous and up-to-date population of the database, the quality control (QC) needed to resolve data conflicts, the synchronization of database tables between unclassified and classified warehouses, and the integration of all data sources into a cohesive database for delivery to the Knowledge Base (KB).

One important challenge in using large data warehouses is the simple and efficient access to the vast holdings within them. Web-based tools have become important assets that address this problem. We have developed web-based tools that enable researchers to do tasks such as track the progress of seismic analysis, access information about stations, origins, and waveforms, view contextual information on a map, handle logistical tasks (i.e., assignment of unique identifiers, track the description and resolution of data problems identified through quality controls), and gain fast access to database metadata (e.g., schema descriptions).

Advances in easy access to metadata are supporting many of the higher-level efforts in quality control, automation, and web access. The first of these is the documentation of the seismic calibration schema using a database schema. This schema is designed to represent all of the detailed table and field information that, up until recently, has been available only in text-based documents. Such information in database form has immediate application to a wide variety of efforts involving the database. (e.g., table creation and quality control, software tools). Another advance in using metadata, with a more narrow application, has been the creation of bulletin descriptive tables. These tables describe the sources of bulletin data that have been imported into the data warehouse, as well as provide a means to track individual data elements to the corresponding lines of text in the original document.

As data become more voluminous and complex, QC has become an increasingly visible and important issue regarding the Knowledge Base. Improvements in QC procedures are helping researchers and data managers to more readily identify complex quality problems. The outcome is consistent research products resulting from improved data upon which those products are based. As we understand the QC problem in more detail, we have begun to automate the process of applying QC to large datasets.

Calibration efforts by Los Alamos National Laboratory (LANL) researchers require working with three separate data warehouses that are physically unable to communicate with each other: two are within LANL; one is located at a remote site. While it is relatively simple to add new data to all warehouses, it is difficult to capture changes made in one and then propagate them to the other two. We have recently developed a procedure based on database triggers to capture these changes. These triggers capture all update, insert, and delete operations against a predefined set of tables. Periodically, the information captured by these triggers is moved to the other environments and executed, thus keeping the warehouses synchronized.

OBJECTIVE(S)

The GNEM R&E program has made recent advances in applying data warehouses to seismic calibration research. Some of the most challenging tasks of maintaining functional data warehouses are the development of software to easily access the contents of the data warehouse, the QC needed to resolve data conflicts, the synchronization of database tables between local and remote warehouses, and the integration of all data sources into a cohesive database for delivery to the (KB). This paper is a brief introduction to the wide range of data management technical issues that we face everyday and the future work needed to fully address all aspects of managing and handling vast amounts of data in a data warehouse that is used in nuclear explosion monitoring research.

RESEARCH ACCOMPLISHED

Web Technology Access to Data Warehouses

As data gathering techniques continue to improve and general data availability increases, the GNEM R&E data warehouses will acquire more data than is readily accessible using the standard SQL command-line interface. One of the challenges is to develop a simple, yet efficient way to view the contents of our data warehouses to assist researchers in developing their calibration products.

Web-based tools provide an efficient, yet easy way to access data from the Oracle GNEM R&E databases. In addition to viewing the seismic data itself, we have developed web pages to interact with database schema viewing and development, handle logistical tasks (i.e., assignment of unique identifiers, tracking of database problems and data requests from researchers, etc.), and view metadata contents (e.g., glossary). The LANL GNEM R&E intranet web technology interface has been operational for over a year and has been extremely useful for accessing data quickly. Recent improvements include being able to generate an origin query, viewing quick, interactive maps of queried data, and implementing a data request system for researchers. Figure 1 is a view of the starting LANL GNEM R&E home page.

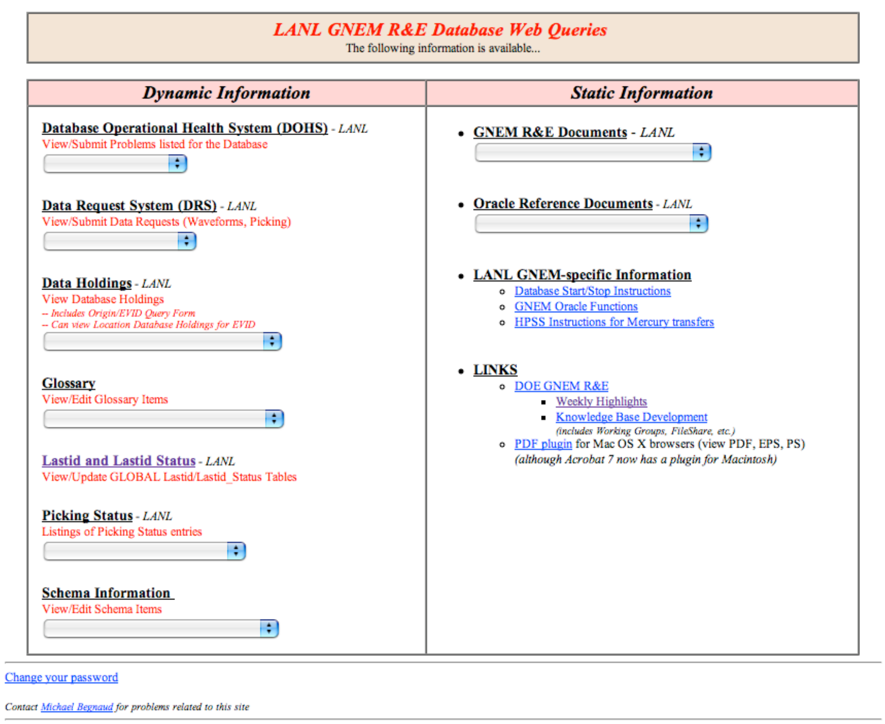


Figure 1. LANL GNEM starting web page. From this page, users can access seismic database entries, glossary and schema information, logistical data, and GNEM-related internal pages.

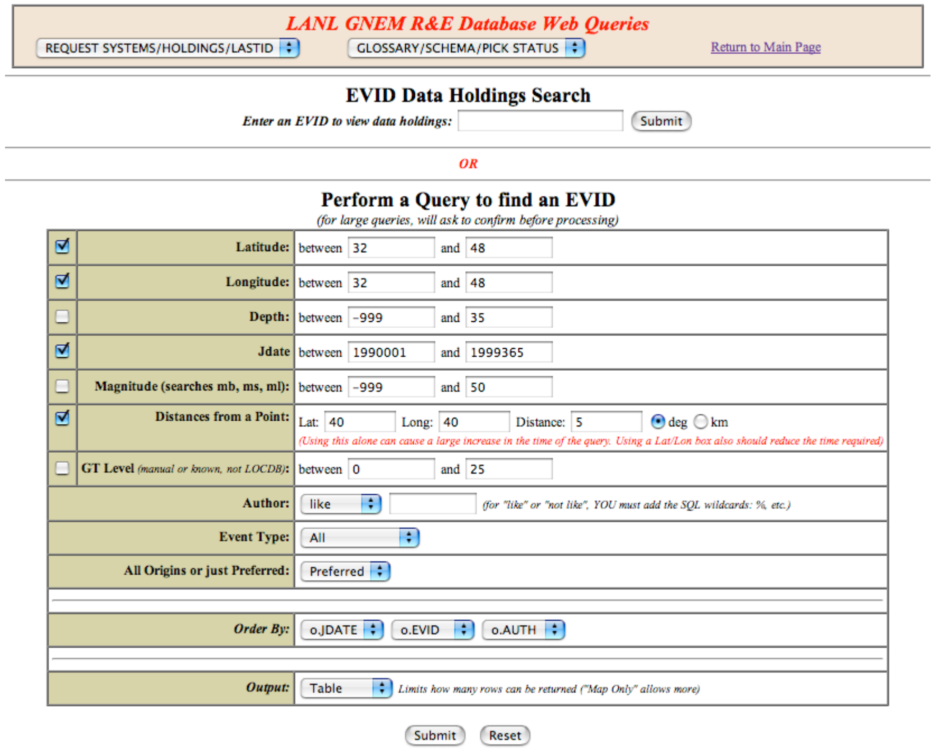


Figure 2. Event query page. Users can enter an EVID directly or perform an origin query. Selected parameters shown were used to generate a query producing results in Figure 3 and Figure 4.

Because of the need for LANL GNEM R&E team members to access data from remote locations, we implemented username/password access as well as Secure Socket Layer (SSL) 128-bit encryption for our internal web technology data access. Within LANL, users are granted access by a browser-standard basic username/password authentication. Outside LANL, users must first use a LANL-authorized cryptocard to gain web access behind the LANL firewall before proceeding to the username/password screen.

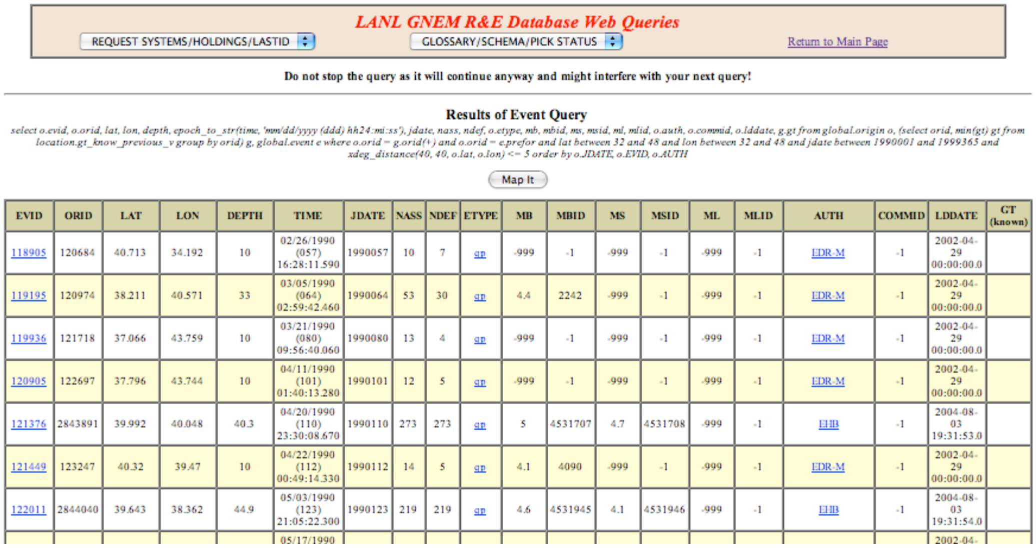
Seismic Data Holdings

The GNEM R&E database schema generally follows the National Nuclear Security Administration (NNSA) structure (Carr, 2005). This structure is mostly centered around events which are built with origins, associations, arrivals, waveforms, etc. Using an Event Identification number (EVID), a user can generally access all the data available for that event. Database users can either enter an EVID directly, or enter parameters for an origin query. Parameters include latitude/longitude, depth, julian date, magnitude, distance from a point, ground-truth value, author or authority, general event type (earthquake, explosion, mining), and all origins or just preferred (Figure 2). Users can request an output HTML table (Figure 3) or just gather data on the web server to produce a map (Figure 4).

In the table output view (Figure 3), a user will see origin information as well as the known ground-truth level. The AUTH and ETYPE fields have cross-referenced links to a glossary table. Clicking one of these links shows the definition of the field information. In the map output view (Figure 4), users can interactively view the results of the origin (or other) queries.

When an EVID is selected from the table output, a new screen appears with an initial summary of available data for that EVID (Figure 5). The web scripts determine if waveform segments, pick status entries, and location database entries exist for that event. Buttons are highlighted for those types available. Since there can be multiple origins for an event, the web page displays each origin and highlights buttons if arrivals, magnitudes, and amplitudes exist for that origin. Users can quickly navigate all data associated with an EVID.

In addition to the EVID-based data holdings, users can query for site-related information, by station abbreviation, reference station, or by entering a manual query (Figure 6). For a single site, the relevant SITE, AFFILIATION, SITECHAN, SENSOR, and INSTRUMENT information are displayed.



LANL GNEM R&E Database Web Queries

REQUEST SYSTEMS/HOLDINGS/LASTID | GLOSSARY/SCHEMA/PICK STATUS | Return to Main Page

Do not stop the query as it will continue anyway and might interfere with your next query!

Results of Event Query

select o.evid, o.orid, lat, lon, depth, epoch_to_str(time, 'mm/dd/yyyy (ddd) hh24:mi:ss'), jdate, nass, ndef, o.etype, mb, mbid, ms, msid, ml, mlid, o.auth, o.commid, o.lddate, g.gt from global.origin o, (select orid, min(gt) gt from location.gt_know_previous_v group by orid) g, global.event e where o.orid = g.orid(+) and o.orid = e.prefor and lat between 32 and 48 and lon between 32 and 48 and jdate between 1990001 and 1999365 and xdeg_distance(40, 40, o.lat, o.lon) <= 5 order by o.JDATE, o.EVID, o.AUTH

Map It

EVID	ORID	LAT	LON	DEPTH	TIME	JDATE	NASS	NDEF	ETYPE	MB	MBID	MS	MSID	ML	MLID	AUTH	COMMID	LDDATE	GT (known)
118905	120684	40.713	34.192	10	02/26/1990 (057) 16:28:11.590	1990057	10	7	qp	-999	-1	-999	-1	-999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
119195	120974	38.211	40.571	33	03/05/1990 (064) 02:59:42.460	1990064	53	30	qp	4.4	2242	-999	-1	-999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
119936	121718	37.066	43.759	10	03/21/1990 (080) 09:56:40.060	1990080	13	4	qp	-999	-1	-999	-1	-999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
120905	122697	37.796	43.744	10	04/11/1990 (101) 01:40:13.280	1990101	12	5	qp	-999	-1	-999	-1	-999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
121376	2843891	39.992	40.048	40.3	04/20/1990 (110) 23:30:08.670	1990110	273	273	qp	5	4531707	4.7	4531708	-999	-1	EHB	-1	2004-08-03 19:31:53.0	
121449	123247	40.32	39.47	10	04/22/1990 (112) 00:49:14.330	1990112	14	5	qp	4.1	4090	-999	-1	-999	-1	EDR-M	-1	2002-04-29 00:00:00.0	
122011	2844040	39.643	38.362	44.9	05/03/1990 (123) 21:05:22.300	1990123	219	219	qp	4.6	4531945	4.1	4531946	-999	-1	EHB	-1	2004-08-03 19:31:54.0	
					05/17/1990													2002-04-	

Figure 3. Partial results of table output for origin query using parameters from Figure 2. Users may select the EVID at left to view specific data (origins, arrivals, netmags, etc.) for that event. Other terms are cross-referenced with glossary tables, giving the definition of the term.

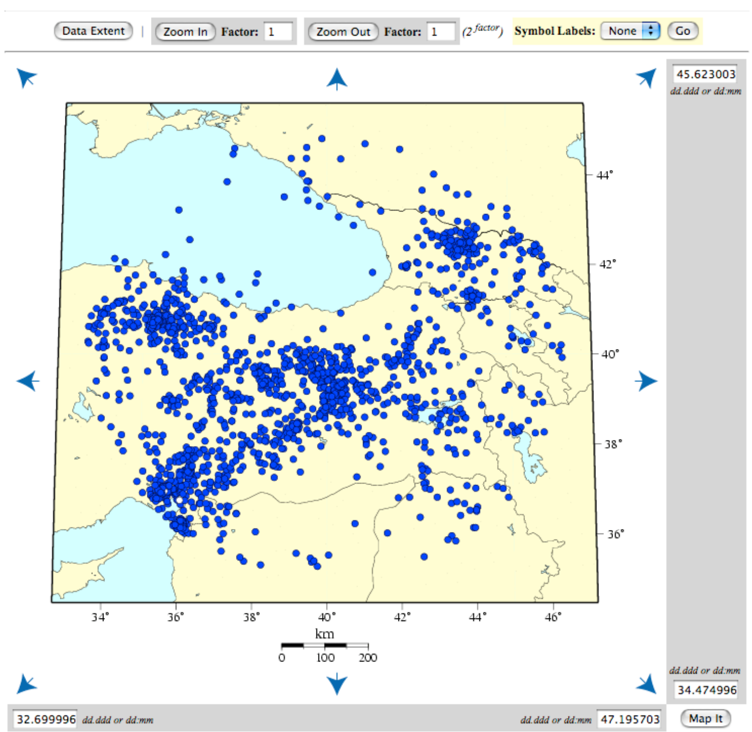
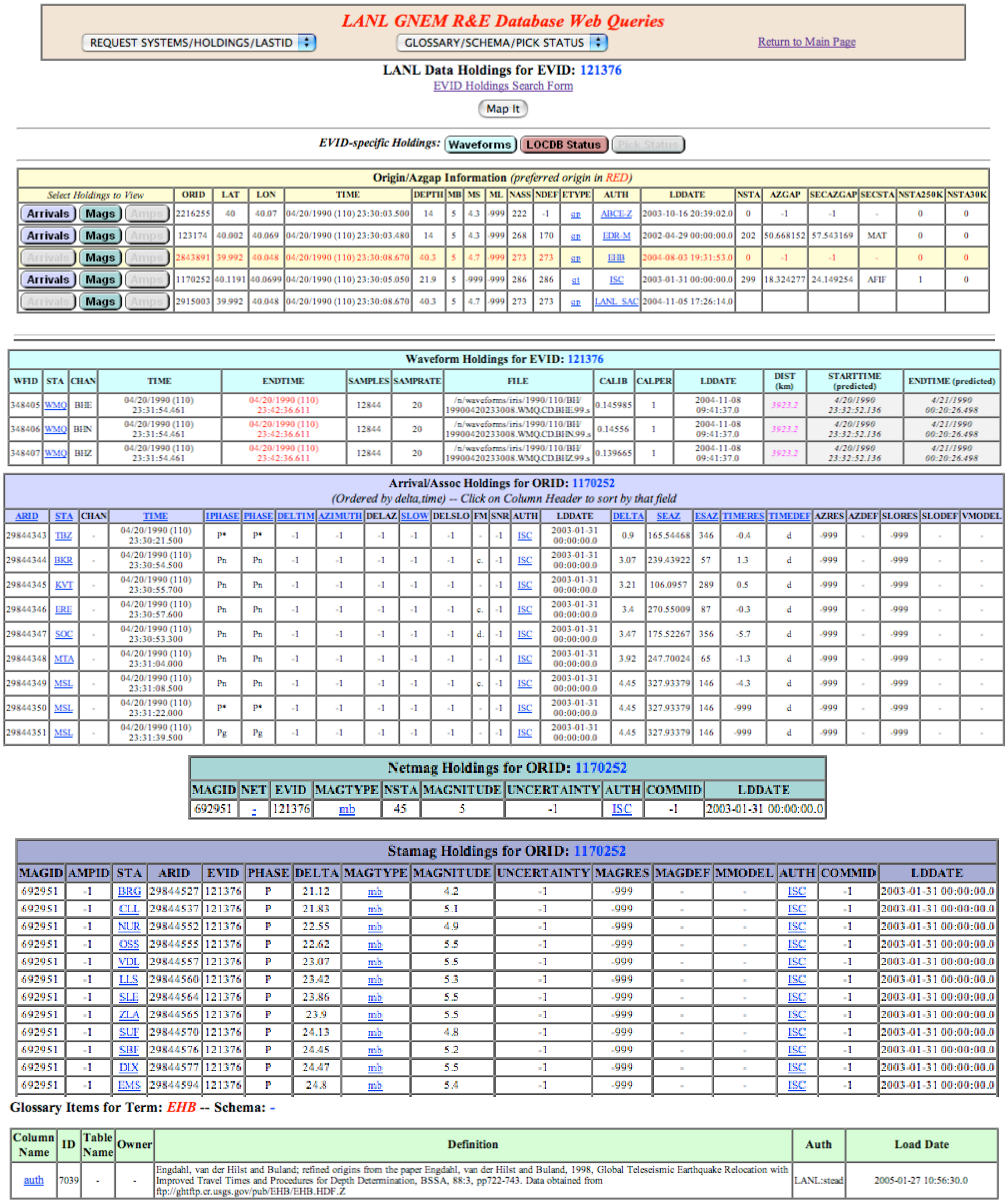


Figure 4. Interactive map of origins produced from query in Figure 2. Users can zoom in or out, set the bounds of the map, and translate the view. Labels can also be turned on or off.



LANL GNEM R&E Database Web Queries

REQUEST SYSTEMS/HOLDINGS/LASTID | GLOSSARY/SCHEMA/PICK STATUS | Return to Main Page

LANL Data Holdings for EVID: 121376

EVID Holdings Search Form

Map It

EVID-specific Holdings: Waveforms | LOCDB Status | Pick Status

Origin/Azgap Information (preferred origin in RED)																					
Select Holdings to View	ORID	LAT	LON	TIME	DEPTH	MB	MS	ML	NASS	NDEF	ETYPE	AUTH	LDDATE	NSTA	AZGAP	SECAZGAP	SECSTA	NSTA250K	NSTA30K
Arrivals Mags Amps	2216255	40	40.07	04/20/1990 (110) 23:30:03.500	14	5	4.3	-999	222	-1	qp	ABCE-Z	2003-10-16 20:39:02.0	0	-1	-1	-	0	0
Arrivals Mags Amps	123174	40.002	40.069	04/20/1990 (110) 23:30:03.480	14	5	4.3	-999	268	170	qp	EDR-M	2002-04-29 00:00:00.0	202	50.668152	57.543169	MAT	0	0
Arrivals Mags Amps	2843891	39.992	40.048	04/20/1990 (110) 23:30:08.670	40.3	5	4.7	-999	273	273	qp	EHB	2004-08-03 19:31:53.0	0	-1	-1	-	0	0
Arrivals Mags Amps	1170252	40.1191	40.0699	04/20/1990 (110) 23:30:05.050	21.9	5	-999	-999	286	286	qt	ISC	2003-01-31 00:00:00.0	299	18.324277	24.149254	AFIF	1	0
Arrivals Mags Amps	2915003	39.992	40.048	04/20/1990 (110) 23:30:08.670	40.3	5	4.7	-999	273	273	qp	LANL_SAC	2004-11-05 17:26:14.0						

Waveform Holdings for EVID: 121376													
WFID	STA	CHAN	TIME	ENDTIME	SAMPLES	SAMPRATE	FILE	CALIB	CALPER	LDDATE	DIST (km)	STARTTIME (predicted)	ENDTIME (predicted)
348405	WMQ	BHE	04/20/1990 (110) 23:31:54.461	04/20/1990 (110) 23:42:36.611	12844	20	/n/waveforms/iris/1990/110/BH/ 19900420233008.WMQ.CD.BHE.99.s	0.145985	1	2004-11-08 09:41:37.0	3923.2	4/20/1990 23:32:52.136	4/21/1990 00:20:26.498
348406	WMQ	BHN	04/20/1990 (110) 23:31:54.461	04/20/1990 (110) 23:42:36.611	12844	20	/n/waveforms/iris/1990/110/BH/ 19900420233008.WMQ.CD.BHN.99.s	0.14556	1	2004-11-08 09:41:37.0	3923.2	4/20/1990 23:32:52.136	4/21/1990 00:20:26.498
348407	WMQ	BHZ	04/20/1990 (110) 23:31:54.461	04/20/1990 (110) 23:42:36.611	12844	20	/n/waveforms/iris/1990/110/BH/ 19900420233008.WMQ.CD.BHZ.99.s	0.139665	1	2004-11-08 09:41:37.0	3923.2	4/20/1990 23:32:52.136	4/21/1990 00:20:26.498

Arrival/Assoc Holdings for ORID: 1170252
(Ordered by delta,time) -- Click on Column Header to sort by that field

ARID	STA	CHAN	TIME	IPHASE	PHASE	DELTIM	AZIMUTH	DELAZ	SLOW	DELSLO	FM	SNR	AUTH	LDDATE	DELTA	SEAZ	ESAZ	TIMERES	TIMEDEF	AZRES	AZDEF	SLORES	SLODEF	VMODEL
29844343	TBZ	-	04/20/1990 (110) 23:30:21.500	P*	P*	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	0.9	165.54468	346	-0.4	d	-999	-	-999	-	-
29844344	BKR	-	04/20/1990 (110) 23:30:54.500	Pn	Pn	-1	-1	-1	-1	-1	c.	-1	ISC	2003-01-31 00:00:00.0	3.07	239.43922	57	1.3	d	-999	-	-999	-	-
29844345	KVT	-	04/20/1990 (110) 23:30:55.700	Pn	Pn	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	3.21	106.0957	289	0.5	d	-999	-	-999	-	-
29844346	ERE	-	04/20/1990 (110) 23:30:57.600	Pn	Pn	-1	-1	-1	-1	-1	c.	-1	ISC	2003-01-31 00:00:00.0	3.4	270.55009	87	-0.3	d	-999	-	-999	-	-
29844347	SOC	-	04/20/1990 (110) 23:30:53.300	Pn	Pn	-1	-1	-1	-1	-1	d.	-1	ISC	2003-01-31 00:00:00.0	3.47	175.52267	356	-5.7	d	-999	-	-999	-	-
29844348	MTA	-	04/20/1990 (110) 23:31:04.000	Pn	Pn	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	3.92	247.70024	65	-1.3	d	-999	-	-999	-	-
29844349	MSL	-	04/20/1990 (110) 23:31:08.500	Pn	Pn	-1	-1	-1	-1	-1	c.	-1	ISC	2003-01-31 00:00:00.0	4.45	327.93379	146	-4.3	d	-999	-	-999	-	-
29844350	MSL	-	04/20/1990 (110) 23:31:22.000	P*	P*	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	4.45	327.93379	146	-999	d	-999	-	-999	-	-
29844351	MSL	-	04/20/1990 (110) 23:31:39.500	Pg	Pg	-1	-1	-1	-1	-1	-	-1	ISC	2003-01-31 00:00:00.0	4.45	327.93379	146	-999	d	-999	-	-999	-	-

Netmag Holdings for ORID: 1170252									
MAGID	NET	EVID	MAGTYPE	NSTA	MAGNITUDE	UNCERTAINTY	AUTH	COMMID	LDDATE
692951	-	121376	mb	45	5	-1	ISC	-1	2003-01-31 00:00:00.0

Stamag Holdings for ORID: 1170252															
MAGID	AMPID	STA	ARID	EVID	PHASE	DELTA	MAGTYPE	MAGNITUDE	UNCERTAINTY	MAGRES	MAGDEF	MMODEL	AUTH	COMMID	LDDATE
692951	-1	BRG	29844527	121376	P	21.12	mb	4.2	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	CLL	29844537	121376	P	21.83	mb	5.1	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	NUR	29844552	121376	P	22.55	mb	4.9	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	OSS	29844555	121376	P	22.62	mb	5.5	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	VDL	29844557	121376	P	23.07	mb	5.5	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	LLS	29844560	121376	P	23.42	mb	5.3	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	SLE	29844564	121376	P	23.86	mb	5.5	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	ZLA	29844565	121376	P	23.9	mb	5.5	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	SUF	29844570	121376	P	24.13	mb	4.8	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	SBF	29844576	121376	P	24.45	mb	5.2	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	DIX	29844577	121376	P	24.47	mb	5.5	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0
692951	-1	EMS	29844594	121376	P	24.8	mb	5.4	-1	-999	-	-	ISC	-1	2003-01-31 00:00:00.0

Glossary Items for Term: ***EHB*** **-- Schema: -**

Column Name	ID	Table Name	Owner	Definition	Auth	Load Date
auth	7039	-	-	Engdahl, van der Hilst and Buland; refined origins from the paper Engdahl, van der Hilst and Buland, 1998, Global Teleseismic Earthquake Relocation with Improved Travel Times and Procedures for Depth Determination, BSSA, 88:3, pp722-743. Data obtained from ftp://ghtftp.cr.usgs.gov/pub/EHB/EHB.HDF.Z	LANL:stead	2005-01-27 10:56:30.0

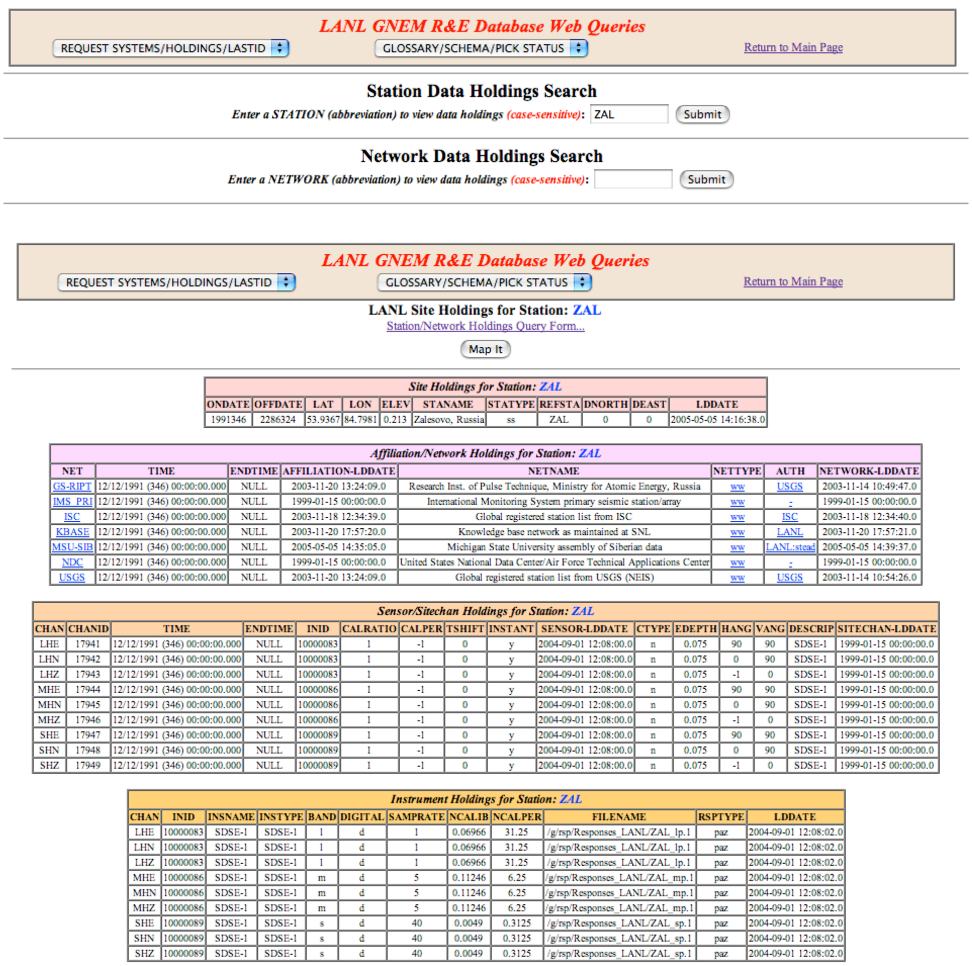
Figure 5. View of EVID-specific data holdings page and several data frames, including a glossary definition. Buttons are highlighted if data are available for that data type. From this page, arrivals, magnitudes (netmag, stamag), and amplitude data can be viewed for the different origins. Users can also view a map of the different origins and stations with waveforms.

Schema Documentation

A major effort in making database metadata available is the documentation of the seismic calibration schema using a database schema. This schema is designed to represent all of the detailed table and field information that, up until recently, has been available only in text-based documents. These documents include versions of the NNSA KB core schema, NNSA KB custom schema, and the United States National Data Center (USNDC) schema documents. The portions that are most needed as readily available metadata are also the portions most amenable to adaptation into database tables themselves: the table descriptions and the column description. Four tables are used to describe schema information: TABDESCRIPT, COLASSOC, COLDESCRIPT, and GLOSSARY. We have also developed a web technology interface to view and edit the schema and glossary information (Figure 7). The schema tables have been accepted as the schema documentation for the GNEM R&E program and are now used by Sandia National Laboratories for complete schema documentation (Carr, 2005). In addition, many of the KB tools developed by Sandia depend on these schema database tables.

TABDESCRIPT provides a basic description of the table, identifies that table with a particular documented schema, and provides a reference in the database to connect the fields that may be associated with the table.

The COLDESCRIPT table not only provides the description of a column but also provides metadata such as NA values, units, and ranges in useful forms. Most numeric ranges have been properly translated into *nmin*, *nminop*, *nmax*, and *nmaxop*. Each operation is relative to the value in the column; that is, 'column *nminop nmin*' and 'column *nmaxop nmax*'. If both are set, then both must apply (implied 'and'). A range type of 'defined' means the value of the column is limited to a short set of predefined values. A 'finite set' is a limited but long or not pre-defined set of values. A 'reference set' is limited to the values in a particular table (as given in *reftab*). Using these fields properly can completely and precisely define the various field ranges.



LANL GNEM R&E Database Web Queries

REQUEST SYSTEMS/HOLDINGS/LASTID | GLOSSARY/SCHEMA/PICK STATUS | Return to Main Page

Station Data Holdings Search

Enter a STATION (abbreviation) to view data holdings (case-sensitive): ZAL Submit

Network Data Holdings Search

Enter a NETWORK (abbreviation) to view data holdings (case-sensitive): Submit

LANL GNEM R&E Database Web Queries

REQUEST SYSTEMS/HOLDINGS/LASTID | GLOSSARY/SCHEMA/PICK STATUS | Return to Main Page

LANL Site Holdings for Station: ZAL

Station/Network Holdings Query Form...

Map It

Site Holdings for Station: ZAL										
ONDATE	OFFDATE	LAT	LON	ELEV	STANAME	STATYPE	REFSTA	DNORTH	DEAST	LDDATE
1991346	2286324	53.9367	84.7981	0.213	Zalesovo, Russia	ss	ZAL	0	0	2005-05-05 14:16:38.0

Affiliation/Network Holdings for Station: ZAL							
NET	TIME	ENDTIME	AFFILIATION-LDDATE	NETNAME	NETTYPE	AUTH	NETWORK-LDDATE
GS-RIPT	12/12/1991 (346) 00:00:00.000	NULL	2003-11-20 13:24:09.0	Research Inst. of Pulse Technique, Ministry for Atomic Energy, Russia	ww	USGS	2003-11-14 10:49:47.0
IMS_PRI	12/12/1991 (346) 00:00:00.000	NULL	1999-01-15 00:00:00.0	International Monitoring System primary seismic station/array	ww	-	1999-01-15 00:00:00.0
ISC	12/12/1991 (346) 00:00:00.000	NULL	2003-11-18 12:34:39.0	Global registered station list from ISC	ww	ISC	2003-11-18 12:34:40.0
KBASE	12/12/1991 (346) 00:00:00.000	NULL	2003-11-20 17:57:20.0	Knowledge base network as maintained at SNL	ww	LANL	2003-11-20 17:57:21.0
MSU-SIB	12/12/1991 (346) 00:00:00.000	NULL	2005-05-05 14:35:05.0	Michigan State University assembly of Siberian data	ww	LANL:stead	2005-05-05 14:39:37.0
NDC	12/12/1991 (346) 00:00:00.000	NULL	1999-01-15 00:00:00.0	United States National Data Center/Air Force Technical Applications Center	ww	-	1999-01-15 00:00:00.0
USGS	12/12/1991 (346) 00:00:00.000	NULL	2003-11-20 13:24:09.0	Global registered station list from USGS (NEIS)	ww	USGS	2003-11-14 10:54:26.0

Sensor/Sitechan Holdings for Station: ZAL															
CHAN	CHANID	TIME	ENDTIME	INID	CALRATIO	CALPER	TSHIFT	INSTANT	SENSOR-LDDATE	CTYPE	EDEPTH	HANG	VANG	DESCRIP	SITECHAN-LDDATE
LHE	17941	12/12/1991 (346) 00:00:00.000	NULL	10000083	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	90	90	SDSE-1	1999-01-15 00:00:00.0
LHN	17942	12/12/1991 (346) 00:00:00.000	NULL	10000083	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	0	90	SDSE-1	1999-01-15 00:00:00.0
LHZ	17943	12/12/1991 (346) 00:00:00.000	NULL	10000083	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	-1	0	SDSE-1	1999-01-15 00:00:00.0
MHE	17944	12/12/1991 (346) 00:00:00.000	NULL	10000086	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	90	90	SDSE-1	1999-01-15 00:00:00.0
MHN	17945	12/12/1991 (346) 00:00:00.000	NULL	10000086	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	0	90	SDSE-1	1999-01-15 00:00:00.0
MHZ	17946	12/12/1991 (346) 00:00:00.000	NULL	10000086	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	-1	0	SDSE-1	1999-01-15 00:00:00.0
SHE	17947	12/12/1991 (346) 00:00:00.000	NULL	10000089	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	90	90	SDSE-1	1999-01-15 00:00:00.0
SHN	17948	12/12/1991 (346) 00:00:00.000	NULL	10000089	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	0	90	SDSE-1	1999-01-15 00:00:00.0
SHZ	17949	12/12/1991 (346) 00:00:00.000	NULL	10000089	1	-1	0	y	2004-09-01 12:08:00.0	n	0.075	-1	0	SDSE-1	1999-01-15 00:00:00.0

Instrument Holdings for Station: ZAL											
CHAN	INID	INSNAME	INSTYPE	BAND	DIGITAL	SAMPRATE	NCALIB	NCALPER	FILENAME	RSPTYPE	LDDATE
LHE	10000083	SDSE-1	SDSE-1	l	d	1	0.06966	31.25	/g/rsp/Responses_LANL/ZAL_lp.1	paz	2004-09-01 12:08:02.0
LHN	10000083	SDSE-1	SDSE-1	l	d	1	0.06966	31.25	/g/rsp/Responses_LANL/ZAL_lp.1	paz	2004-09-01 12:08:02.0
LHZ	10000083	SDSE-1	SDSE-1	l	d	1	0.06966	31.25	/g/rsp/Responses_LANL/ZAL_lp.1	paz	2004-09-01 12:08:02.0
MHE	10000086	SDSE-1	SDSE-1	m	d	5	0.11246	6.25	/g/rsp/Responses_LANL/ZAL_mp.1	paz	2004-09-01 12:08:02.0
MHN	10000086	SDSE-1	SDSE-1	m	d	5	0.11246	6.25	/g/rsp/Responses_LANL/ZAL_mp.1	paz	2004-09-01 12:08:02.0
MHZ	10000086	SDSE-1	SDSE-1	m	d	5	0.11246	6.25	/g/rsp/Responses_LANL/ZAL_mp.1	paz	2004-09-01 12:08:02.0
SHE	10000089	SDSE-1	SDSE-1	s	d	40	0.0049	0.3125	/g/rsp/Responses_LANL/ZAL_sp.1	paz	2004-09-01 12:08:02.0
SHN	10000089	SDSE-1	SDSE-1	s	d	40	0.0049	0.3125	/g/rsp/Responses_LANL/ZAL_sp.1	paz	2004-09-01 12:08:02.0
SHZ	10000089	SDSE-1	SDSE-1	s	d	40	0.0049	0.3125	/g/rsp/Responses_LANL/ZAL_sp.1	paz	2004-09-01 12:08:02.0

Figure 6. Site query and information pages for a single station (ZAL). In addition to standard site information, affiliation, sitechan, sensor, and instrument data are also displayed. Users may also plot a map of the site.

The COLASSOC table associates a given COLDESCRIPT with a TABDESCRIPT. The table will have multiple columns (column positions 1 through the total number, in order), and columns can appear in multiple tables. The column type will define the basic function of the field in the database (i.e., primary key, unique key, descriptive data, measurement data, administrative data). *Key* allows key columns to be identified with respect to the table: the reference table for the key, or a table in which the key is foreign. These keys are primarily numerical identifiers. Keyschema is used when the reference table is part of a separate schema.

The GLOSSARY table serves two purposes: The first is to simply define generic strings used in various description fields, primarily acronyms and abbreviations. The second is to serve as the reference table for 'defined' range types and 'finite set' range types. A given definition can apply in all circumstances (column_name, table_name, owner and schema are all not set), which is a generic definition, or it can apply to increasingly selective subsets of columns, tables, owners, and schemas. For 'defined' and 'reference set' range types in COLDESCRIPT, the complete permitted set of values will be found in GLOSSARY, one entry for each value (values are found in the name column), and *column_name* will always be set for these.

Having such schema information in database form has proven beneficial to a wide variety of efforts involving the database, both specific to Los Alamos, and across NNSA. This schema for table description metadata is a key contribution to the web-based database documentation discussed in this paper. It is also the foundation for automated QC efforts at LANL (see below). The schema tables have changed slightly over the past year, in response to experience using it at LANL and elsewhere. The tables provide immediate advantages in the maintenance of the schema descriptions, such that they can be easily checked for errors, quality, and completeness. The use of the TABDESCRIPT, COLDESCRIPT, and COLASSOC tables is becoming fairly well established. The GLOSSARY table has more recently seen greater use and has proven helpful in QC of text fields and in finding full descriptions of various text-based values such as phase names or authors (e.g., "just what is SPdifKS?", "is there a reference for "SIB:AT62"?").

<u>*Logistical Information*</u>

The LANL web technology access not only displays seismic and schema information, but has interfaces to allow researchers to handle logistical tasks such as requesting waveform, picking, and catalog data, submitting problems encountered with the database, and viewing identifier values for database primary key fields. Both of these functions

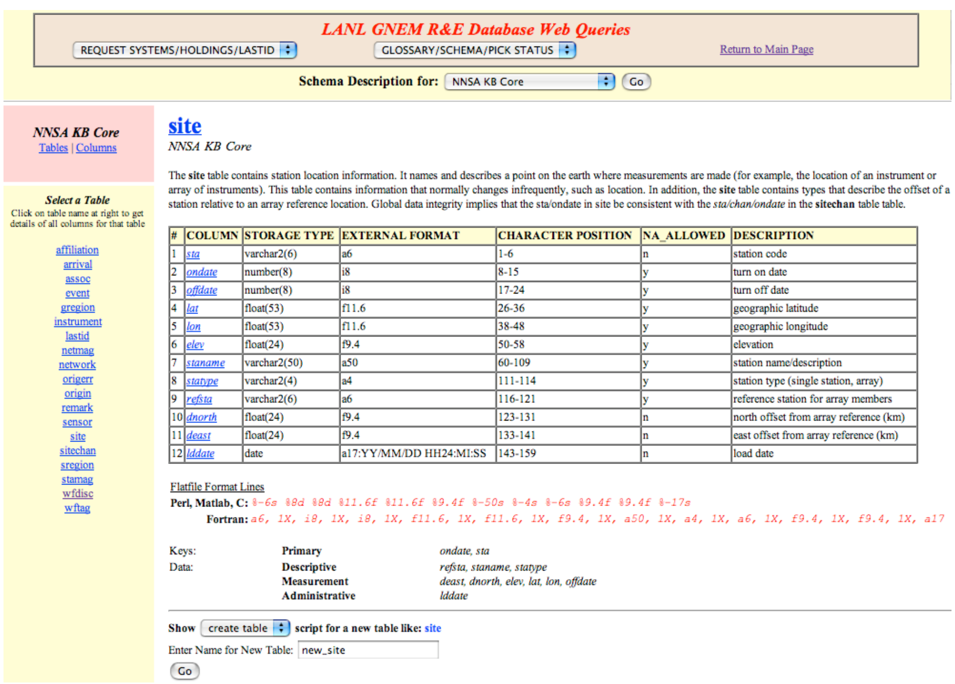


Figure 7. View of schema web page. Any schema can be displayed with links pointing to table descriptions as well as individual column descriptions. Other web pages are also used to directly edit the schema information.

enable tracking of pending and completed data requests and listed database problems.

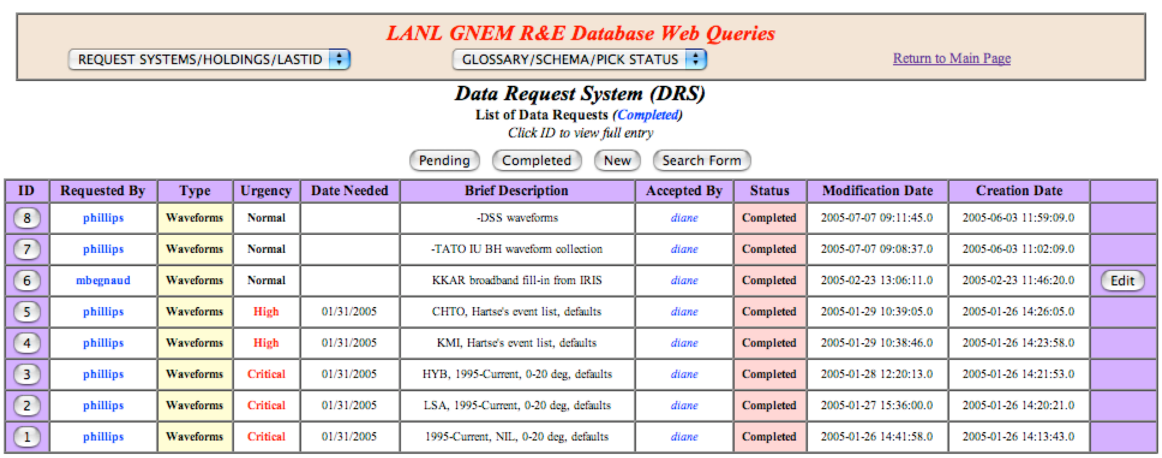
For the new Data Request System (DRS) (Figure 8), users can request that waveform, arrival picks, or catalog data be acquired for use in their research projects. Requests are entered into the system, logging who made the request along with the request details, and notifications are automatically sent to members of the LANL Data Management Team. A member of the team then accepts the request, retrieves the data, and sets the status of the request to "Completed." During this process, the person who originally made the data request is automatically notified of changes in status or comments from the "Acceptor." This formal method for data requests allows tracking of researcher needs and reduces instances of miscommunication.

Quality Control and Assurance

As data become more voluminous and complex, QC has become a challenging and extremely interesting issue to consider when developing content for the KB. In particular, because of recent advances in tools that access KB data, researchers have been able to make more efficient use of this large volume of data but have also found inconsistencies in applying the data to their research efforts. Improvements in QC procedures are helping researchers and data managers to more readily identify complex quality problems. The outcome is improved research products resulting from improved data upon which those products are based. QC is handled in a wide variety of ways at the present, and much effort is being made to better structure this procedure and automate as much of it as is practical to do so.

There are three main categories of QC: manual, tool-assisted, and automated. Improvements in manual QC are occurring constantly as a wider variety of issues are captured and understood. But this kind of QC is the least transportable and repeatable. The next step is to capture the tracking and resolution of QC problems in various simple tools, usually case-specific scripts. Tool-assisted QC is more transportable and repeatable, since the script serves as documentation of procedures but it still requires case-by-case modification and application. Automated QC is preferred and cannot happen without there first being a fairly comprehensive understanding of the problem.

The best approach to automated QC is to document exactly what the database should be and reject anything that does not conform. This approach is why the database descriptive schema discussed above has been a valuable tool in automated QC. This cannot address all QC issues (for example, QC of waveforms), but will handle the bulk of the information in the database. LANL now has an automated process that can be configured to run a very comprehensive QC against a wide variety of data sets that may be incorporated into the KB. Since it is based on the content of the schema tables (i.e., TABDESCRIPT, COLASSOC, COLDESCRIPT, GLOSSARY), it can readily handle the addition of new custom tables for particular data sets. It produces a comprehensive QC report that greatly speeds the identification of problems that need to be addressed. This was of great help, for example, in preparing the recently delivered Siberian dataset, which was a highly heterogeneous collection of data from a wide variety of sources. The process is based on the schema tables, uses a simple parameter file, and implements single-column and single-table tests, two-table joins, and the notorious "wftag"-type join. It also extends these tests and joins using a special database table called COMPLEXJOIN, that permits a wide variety of complex relationships including


LANL GNEM R&E Database Web Queries

REQUEST SYSTEMS/HOLDINGS/LASTID | GLOSSARY/SCHEMA/PICK STATUS | Return to Main Page

Data Request System (DRS)

List of Data Requests *(Completed)*

Click ID to view full entry

Pending | Completed | New | Search Form

ID	Requested By	Type	Urgency	Date Needed	Brief Description	Accepted By	Status	Modification Date	Creation Date	
8	phillips	Waveforms	Normal		-DSS waveforms	diane	Completed	2005-07-07 09:11:45.0	2005-06-03 11:59:09.0	
7	phillips	Waveforms	Normal		-TATO IU BH waveform collection	diane	Completed	2005-07-07 09:08:37.0	2005-06-03 11:02:09.0	
6	mbegnaud	Waveforms	Normal		KKAR broadband fill-in from IRIS	diane	Completed	2005-02-23 13:06:11.0	2005-02-23 11:46:20.0	Edit
5	phillips	Waveforms	High	01/31/2005	CHTO, Hartse's event list, defaults	diane	Completed	2005-01-29 10:39:05.0	2005-01-26 14:26:05.0	
4	phillips	Waveforms	High	01/31/2005	KMI, Hartse's event list, defaults	diane	Completed	2005-01-29 10:38:46.0	2005-01-26 14:23:58.0	
3	phillips	Waveforms	Critical	01/31/2005	HYB, 1995-Current, 0-20 deg, defaults	diane	Completed	2005-01-28 12:20:13.0	2005-01-26 14:21:53.0	
2	phillips	Waveforms	Critical	01/31/2005	LSA, 1995-Current, 0-20 deg, defaults	diane	Completed	2005-01-27 15:36:00.0	2005-01-26 14:20:21.0	
1	phillips	Waveforms	Critical	01/31/2005	1995-Current, NIL, 0-20 deg, defaults	diane	Completed	2005-01-26 14:41:58.0	2005-01-26 14:13:43.0	

Figure 8. Data Request System page showing "Completed" requests.

grouping relationships (i.e., comparing origin.nass to assoc), multi-table joins (ex: comparing wfdisc.instype to instrument.instype), and computations (i.e., comparing origin.time to origin.jdate). Future development in this area will examine the possibility of automating the repairs, in addition to just the QC. In general, QC remains a large problem and GNEM R&E is making new and unique contributions toward resolving this important problem.

Bulletin Descriptive Tables

An advance in using schema information for QC has been the creation of bulletin descriptive tables. These tables describe the sources of bulletin data that have been imported into the data warehouse, as well as providing a means to track individual data elements to the corresponding lines of text in the original document. There are two tables involved: BULLETIN and BULLASSOC. The BULLETIN table contains one entry for each individual bulletin, with columns *dir* and *dfile* pointing to the text file on the system that contains the bulletin. It also has a *bullid* column that is unique to each bulletin. The other information in the table describes the bulletin, including format. The BULLASSOC table has one line for each data object extracted from the bulletin (origins, arrivals, magnitudes, etc.) It links the *bullid* and the *id* of the extracted object and provides the line number in the bulletin corresponding to the object. These tables have immediate use in QC. First, they allow problematic objects to be traced directly to the corresponding file and line number. Second, they can be used to extract the entire contents of a single bulletin from the integrated database when the need to remove or replace the data from a particular bulletin arises.

Segmenting Continuous Waveforms

Over the years LANL has acquired segmented and continuous waveforms from many different sources in formats such as SEED, SAC, CSS (with accompanying WFDISC lines), GSE, and SEGY. To make these data readily available to researchers, we have developed a PERL code that uses a database interface (the “PERL DBI”) to assemble user-specified, event-based wave segments into SAC files. We call this code "wfdisc2sac.pl", because it requires a database WFDISC line description of each waveform that might be cut and transformed into SAC format.

For the case of SEED data handling, wfdisc2sac.pl calls the executable “rdseed”. To run the code, a user builds a list of EVIDs that correspond to events of interest and specifies a list of desired stations and channels. The user can also specify desired time window lengths of the final SAC waves based on Jeffreys-Bullen travel-time tables. If ORIGIN, SITE, and ARRIVAL tables are available, wfdisc2sac.pl will use the PERL DBI to query these tables and find information to populate the newly created SAC header fields. To prevent accidental recutting of segments already listed in the LANL WFDISC table, an option is available to check for existing segments prior to attempting a fresh cut on continuous data. A second PERL code builds WFDISC flat file lines that can be immediately inserted into the WFDISC table.

Database Synchronization - Capturing and Propagating Data Changes

Calibration efforts by LANL researchers require the use of three separate databases that are physically unable to communicate with each other: two within LANL, and one at a remote site. Because of the lack of direct communication between these databases, maintaining data synchronization between them is difficult. The content of these databases is such that some data are common to all the databases, some data are common between only two of the databases, while some data are allowed to exist only at the remote location. While it is relatively simple to add new data to all databases, it is difficult to capture changes such as updates or deletes made in one and then propagate them to the other two.

We have recently developed a procedure based on database triggers to capture changes made to core database tables. This procedure has been in place for about one year, and results to date have been satisfactory. The changes being monitored on the predefined set of tables are data inserts, updates, and deletes. This process is referred to as Capture Data Changes (CDC). The acronym, as well as the fundamental idea, is similar to Oracle’s Change Data Capture method of implementing the incremental recording of data changes. The main difference between Oracle’s implementation and LANL’s is that our method does not depend on a particular version of the Oracle Relational Database Management System (RDBMS). Oracle’s CDC method is directly tied to a specific application that must be installed, configured, and run against an Oracle 9i database. Our procedure is entirely based on database triggers, which are available on any version of the Oracle RDBMS; thus, our implementation is not tied to any particular version of the Oracle database and can be implemented on any platform.

The concept is simple. Database triggers are created against a predefined set of tables to be monitored. These triggers fire upon insert, update, or delete operations against these tables. The triggers capture unique information about a row being inserted, deleted, or updated. Our implementation of capturing only the information needed to uniquely identify a changed row in a table leads to significant disk space savings.

An interesting by-product of our synchronization procedure is that we not only capture unique information about the rows being modified in the tables being monitored, but we also capture the modification date and the name of the database user who made the modification to each row. This information, together with the built-in database auditing capabilities that can be enabled at the table level, can serve the secondary purpose of providing a security audit trail to be able to answer questions regarding changes made to critical production data.

The synchronization operation between the source database and the two target databases is a manual process at this time. The synchronization operation starts after a predetermined number of changes have occurred in the source database tables. A special set of tables is created from the unique information captured by the triggers that contain the table structure and changed rows of data from the source tables that will be used to replace the outdated information in the target databases. The second and final step of the synchronization process is done on the target databases, both at LANL and at the remote location. First, we delete from and insert rows into the target tables. The rows to be deleted in the target databases are the rows that were either updated or deleted from the original source tables. In this step, the decision was made to replace the entire row when an update occurred in the source table, rather than try to make a column-by-column comparison to only update specific columns in the target table. The latter choice would be costlier in terms of computer resources.

CONCLUSIONS AND RECOMMENDATIONS

Developing web technology interfaces to handle common database queries has allowed LANL researchers to access data more readily and efficiently. Specific information for seismic events can be retrieved quickly with ties to relevant information. Utilizing protected web interfaces allows users to view data remotely and helps in synchronizing data retrieval and processing tasks. We are continually finding new reasons and ways to securely access the database through the web. Future web technology development includes tracking processing steps for waveforms, picking, amplitudes, etc., as well as improving QC checks. In addition, we are working on better methods for viewing and interacting with waveform and map data via a web interface.

QC has become an increasingly visible and important issue regarding the KB, as data have become more voluminous and complex. Improvements in QC procedures will help researchers and data managers to more readily identify complex quality problems. The outcome is improved research products resulting from improved data upon which those products are based.

The process of CDC shows great promise. It is expected that we have not encountered all possible use cases, and modifications to the process will need to be made and perhaps auxiliary tables or triggers will need to be created. This process has been in place for about one year and it appears to be successful. Many of the steps involved in synchronizing local and remote databases are manual, and direct interaction with the databases is needed throughout the process. Future work includes the automation of many of the synchronization steps and the incorporation of appropriate quality control checks to ensure that the synchronization between databases was successful.

ACKNOWLEDGEMENTS

The authors wish to acknowledge LANL personnel who have made important contributions in the past and those who continue to make vital contributions to the development and maintenance of the LANL research data warehouse: Diane F. Baker, Marian D. Peters, W. Scott Phillips, George E. Randall, James T. Rutledge, and Steven R. Taylor,

REFERENCES

Carr, D. (2005), *National Nuclear Security Administration Knowledge Base Core Table Schema Document*, Sandia National Laboratories report SAND2002-3055.