



code	stationID/ sub	ship name	cruise no.	cruise day	cruise month	cruise year	Local time	time zone	general area
100	A002	AB3	70	7	12	55	200	5	NANTUCKET SHOALS
100	A003	AB3	70	7	12	55	510	5	SW CHANNEL
100	A020	AB3	70	11	12	55	1850	5	SW GULF OF MAINE
100	E001	AT	263	25	5	61			SHELF S OF MARTHAS VN
100	E002	AT	277	23	5	62			SLOPE S OF NEW ENGLND
100	2278	AST	3	24	5	65	1142	5	ALTAMAHA RIVER,GA.
100	2279	AST	3	25	5	65	555	5	ST.SIMONS SOUND,GA.

Figure 1 Structure of a USGS text file (one example of 20 formats for a station)

code	sampleID/ sub	%CaCO3	type microscope	number of mollusks	number of echinoids	number of benthics	number of planktonics	number of bryzoans	number of serpulid worms	number of barnacles	number of algae
310	1484	33	B	75	1				5	1	5
310	1485	18	B	65	5		2		1	3	2
310	1486 A	12	B		F	C		C			
310	1486 B	8	B		F	C		C			
310	1487	17	B	76			5		1	3	11
310	1488 A	7	B		F	C		C			
310	1488 B	4	B		F	C					R
310	1489	6	B	76	6		2			2	10
310	1490	13	B	77	5				2	1	7
310	1491	8	B	80	11					-1	8

Figure 2 Example of data definition variation in a column (USGS text file)

Scientists around the world are continuously gathering large amounts of geological data; it was from these sources that we obtained our sample data. Because SQL Server is easier to use than traditional statistical methodology, we decided to pool existing data sources into a SQL Server data warehouse. We also had to use SQL Server to verify the original analyses of the geological data, which were performed through traditional statistical methods. We built two data warehouses from two different geological data sources. These sources were comparatively small because we were doing an academic project, but we believe that the findings would hold true no matter the size of the data set. One data set consisting of more than 20 files was a study of the Eastern Continental Shelf that had been collected and analyzed by the United States Geological Survey (USGS) between 1955 and 1970. The second database contained the results from a study of the decomposition effects of burial on six different wood species in the Bahamas and Gulf of Mexico. The amount of wood standing or decaying in water throughout the world is phenomenal. The decomposition of wood produces greenhouse gases, but this process stops when the wood is buried. This study is significant because it provides scientists additional valuable information about the sources of hydrocarbon emissions into our atmosphere.

Cleaning the Original USGS Data

The USGS results in a data set consisting of 20 text files, including both analytical and descriptive data types, was previously published by the National Geophysics Data Center (NGDC). (See Hathaway, John C., 1971 Data File, Continental Margin Program, Atlantic Coast of the United States: WHOI Reference No. 71-15.) We created a composite primary key based on existing columns that held the station number, sample ID, and sub-sample letters. And we set the granularity at the sub-sample (the sub-sample code identifies the samples that were divided from a larger sample). Figure 1 shows sample data from one of the files.

To clean the original USGS data, we imported it first into Microsoft Excel, then into Microsoft Access, and finally into SQL Server 2000. We took this circuitous route because of the large variation in data format and definition within the text files. The import functions in SQL Server and Access weren't as robust as that of Excel. Using the Excel Text Import Wizard, we manually set

the field widths and formats for each col-



umn, creating new columns for amalgamated data and splitting data fields. This process eliminated much of the data variation. We then loaded the data into Access, using Access's *import external data* option. During the import, Access found more data-formatting errors; it stored the row numbers of incompatible data in an import error file for reference. Figure 2 shows one problem we encountered: Data formats within two columns changed from numeric to alphabetic, then back to numeric. To correct the errors, we used a combination of editing techniques in Excel and wrote SQL code to reformat the data within the Access table. We then manually edited some of the data and used UPDATE queries to change other data to its numerical equivalent.

We couldn't import some files into Access because of data-type incompatibility errors in the Sample ID field. This field is alphanumeric, but Access was trying to format it as an integer. We tried to use the Excel Text Import Wizard to modify the

site	depth (m)	yr	mag	quer	arau	seq	pinus	q. stell
AA50RID	15	2	60	90	80	90	50	90
AA100S1	30	2	95	90	90	85	95	60
AA240LE	73	2	90	80	90	50	90	70
AA290WA	88	2	60	60	60	60	90	95

Figure 3 Sample data from the wood study

area zone	sheet no.	navigation method	Latitude degrees	Latitude minutes	Longitude degrees	Longitude minutes	corrected depth	sounding device
18	1	2	41	6	69	17	51	1
17	1	2	40	51	68	55	66	1
13	1	2	41	54	69	37	185	1
23	1	2	40	20.5	70	47	97	1
24	1	2	39	54.5	70	35	487	1
51	3	1	31	19.98	81	23	2 4	1
51	3	1	31	7.93	81	24.8	3 8	1

number of halimeda	number of coral	number of pelletoids	number of ooids	number of ancrustation	number of encrusted/ altered	number of lithoclasts	number of unknown carbonate	number of misc.	assemblage code
2					7				M
2			7		11				M
				C					M
				C					M
1					5				M
									M
									M
					4				M
2				-1	6				M
								M	

data format; however, when we imported the data into Access, Access didn't recognize the Excel character format. We then tried to design the table in Access and import external data, but Access still overrode the character format. As a workaround, we loaded the data into SQL Server, declaring the field as a character type, then re-imported it into Access for field identification and data cleanup. We found table manipulations easier to manage in Access 2000 than in SQL Server 2000.

Cleaning Up the Wood Study Data

Figure 3 shows the data from the wood study, which had been published in the dissertation of Elizabeth Heise (*A comparison of the fossilization potential and recycling of wood by wood-boring bivalves and isopods on the shelf and slope of the Bahamas and Gulf of Mexico*, Texas A&M University, 2001) and was therefore quite clean. The most difficult task was dealing with hidden nulls in a column with a text data type that we were using for a primary key field. The null fields were invisible, so we didn't know they were a problem until Data Transforma-

tion Services (DTS) refused to transform the Excel spreadsheet into a SQL Server table. The null fields were invisible in the Excel worksheet, and they didn't appear when the tables were printed. We removed the primary key designation from that column so that the DTS Import Wizard would allow nulls in it. Then we used DTS to create the table in SQL Server format. When we looked at the table with all rows open in SQL Server 2000, three rows of nulls showed up at the bottom of the SQL Server table. As an afterthought, we used the DTS preview option to view the files that contained the bad records. The nulls were then visible. Based on this experience, we

recommend using the DTS preview option on all files before attempting to load them into SQL Server.

We edited the Excel spreadsheet, pasting into a new Excel spreadsheet only the rows that contained data. We deleted all the data from the SQL Server table and ran the DTS Import Wizard again to produce eight relational tables in SQL Server format. Figure 4 shows the five tables that represent one of the two locations in the wood study.

Building the Data Warehouses

For the geological data warehouse, we used

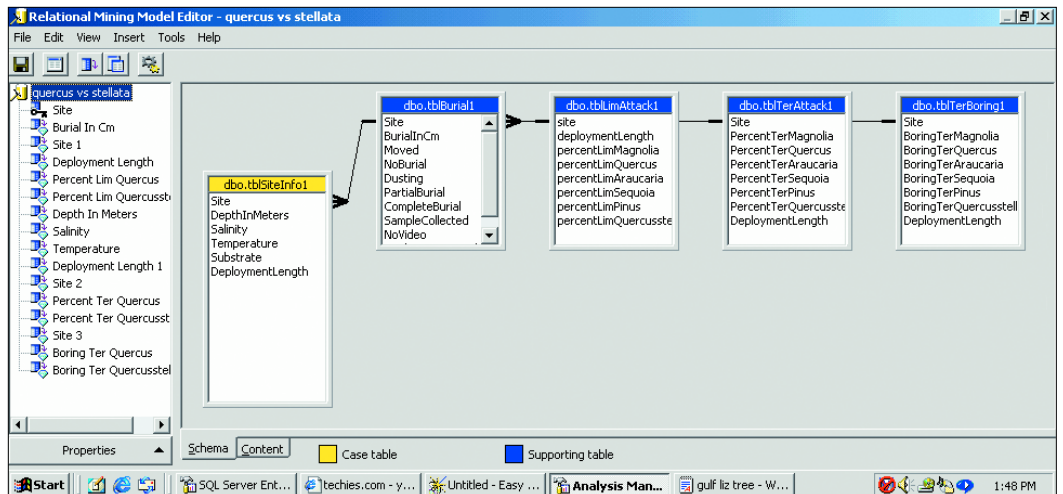


Figure 4 SQL Server tables containing the geological data