RELATIONAL DATABASES AS A TOOL TO MANAGE ENVIRONMENTAL DATA AT THE RESEARCH PLOT SCALE

Tyler Erickson[1] and Deborah Donahue[2]

ABSTRACT

Relational databases management systems (RDBMS) can be used as highly effective storage systems environmental data. RDBMS are most suited to storing large volumes of repetitive data. RDBMS offer significant advantages over data storage methods (such as ASCII text files or spreadsheet programs) by offering flexibility in the types of information stored, speed of retrieval, and the ability to share data with other computer programs. Information contained in RDBMS can be accessed through desktop programs or may be distributed over the Internet. However, RDBMS may not be appropriate for all types of collected data, especially when the format of the data changes frequently.

Snow hydrology data can be efficiently stored in a RDBMS, particularly if many repetitive measurements are made in time or in space, such as automated meteorological measurements obtained from data loggers. Intelligently designed RDBMS link metadata to the collected data and allow scientists to analyze and store comments on the quality of the data. Design issues encountered during the processes of data modeling, database design, and data distribution are described. Examples are taken from the currently operational Mammoth Mountain Energy Balance Monitoring Site database and a sample data model created to store information on Western Snow Conference (WSC) members and papers.

INTRODUCTION

Snow hydrologists collect a large variety of data. The initial data is often recorded on standardized datasheets, by automated data loggers, and in field notebooks. Unless the hydrologist is the sole user of the data, the data needs to be transferred to another system that allows for the distribution of the data to multiple users. The sharing of collected data benefits all snow hydrologists, because high-quality data sets are difficult and expensive to collect. Many storage solutions exist, each with its own strengths and weaknesses. Relational database systems can provide a good solution for distributing snow data sets, particularly for large volumes of data that has a constant format.

This paper provides an overview of relational database systems and their terminology. Individual topics, such as data modeling, have entire textbooks devoted to covering them, and area therefore not discussed in detail here. Several currently operational web-accessible database systems are discussed and an example data model that could be used to store information on members of the Western Snow Conference and the papers presented at the conference is presented.

THE NEED FOR RELATIONAL DATABASES

There are many options available for storing environmental data. A few common options are: hardcopy notes, text data files, spreadsheets, relational databases, and GIS systems. These options are roughly ordered by increasing difficulty to set up, flexibility of distribution, ability to handle large volumes of data, and form unstructured to structured data.

Hardcopy Notes

Field observations are commonly recorded in field notebooks or standardized datasheets. Later the data can be transferred to a computer format. Advantages of hardcopy notes are that the format is easy to use in the field and the format is flexible. Disadvantages of hardcopy notes are that they are not easily shared with other users, they can take a significant amount of physical space in storage, and that the format becomes impractical for large datasets.

Text Data Files

Text data files can be created from hardcopy notes or can be obtained from automated data loggers. Advantages of text files are that almost any format of data can be stored and that text files are easy to distribute among other users via a local network or the Internet. Furthermore, all computer users are generally comfortable with opening and using text files. The primary disadvantage of text files is the lack of flexibility in searching the

[1] Institute of Arctic and Alpine Research and Department of Geography, University of Colorado at Boulder, 1560 30th Street, UCB 450, Boulder, CO 80309, tyler.erickson@colorado.edu

[2] Data Manager, Snow Hydrology Research Group, Donald Bren School of Environmental Science and Management, University of California

data. The flexibility of the format can be a disadvantage, by potentially allowing for the storage of incorrectly formatted data.

Spreadsheets

Advantages of storing data in a spreadsheet are that preliminary data analyses (equations, graphs, etc.) can be stored along with the data and that the data is easily distributable over the Internet. Disadvantages are that they are inefficient in terms of file size, there is often a finite limit to the amount of data that can be recorded, distributing the files often leads to the existence of multiple similar copies, and spreadsheet formats are specific to a given computer operating system.

Relational Databases

Relational databases allow for the efficient storage of repetitive data. Although the term 'database' may also apply to non-relational and object-relational databases, in this paper it refers to the most common type of database: to relational database.

To design a database, the format of the data has to be known in advance, and the database tables must be planned and created prior to data entry. Databases can be set up to limit the type and range of data that is entered which can limit the occurrence of data errors. Databases are efficient users of disk space and the amount of data that can be stored is generally limited only by the physical storage device. Multiple users can simultaneously enter and retrieve data from a database within a local network or across the Internet.

Disadvantages of databases are that they are work-intensive to set up and maintain and that some specific knowledge is required to extract data. Databases are not well suited to storing unstructured data because the database tables need to be created before data is imported.

Geographic Information Systems (GIS) are used to store and analyze data that have spatial attributes. Modern GIS applications store their geographic data in an underlying relational database, and the data can be accessed from outside the GIS application. Because GIS use a database for data storage, they will not be discussed further in this paper.

DATA MODELING

Designing a database involves a process known as 'data modeling' in which the designer identifies the unique 'entities' which are to be stored and the relationships between the entities. An entity is a person, place, thing, or concept that has characteristics about which you want to store information.

A relationship is an association between entities. There are several types of relationships, which depend on the cardinality of the association and whether or not the entities are required. Cardinality is the numeric relationship between occurrences of entities. Some examples of relationships with different cardinality are:

One-to-One Relationship Example: each Western Snow Conference (WSC) member has a single social security number and each social security number can be used by only one WSC member.
One-to-Many Relationship Example: each WSC member lives in a single region (state or province) but each region may contain the home(s) of one or multiple WSC members.
Many-to-Many Relationship Example: each WSC member may be an author on 0,1, or multiple papers presented at the conference and each paper may have 1 or multiple authors

Entities often have multiple relationships with other entities. A figure that shows the organization of entities and relationships is called an Entity-Relationship (ER) diagram. An example ER diagram for the WSC meeting is shown in Figure 1. The data model allows for the storage of information about the WSC members and the papers presented at the meeting. Each entity has a distinct purpose:

- Member: Contains information on WSC members. Sample data for this entity are shown in Figure 2.
- Paper: Contains information on papers presented at the WSC meeting.
- Member to Paper: Links WSC members to papers.
- Member Type: Contains information on different classes of WSC members. Sample data for this entity are shown in Figure 3.
- Country: Contains information on countries. Sample data for this entity are shown in Figure 4.
- Region: Contains information on regions within countries. Sample data for this entity are shown in Figure 5.

Database tables are created based on the entities defined in the ER diagram. Each table has a number of columns that are available to store data. Each column is defined with a particular data type (integer, decimal,

character, etc.) and is designated as a required or optional column. An example of a complex system of tables is shown in Figure 6.

The process of data modeling and creating database tables forces the uses to consider the format and purpose of the data prior to storage. This process can slow down the initial implementation of a database system, yet it results in a final product that is well planned.

A good overview of data modeling is given by the University of Texas at Austin website [5].

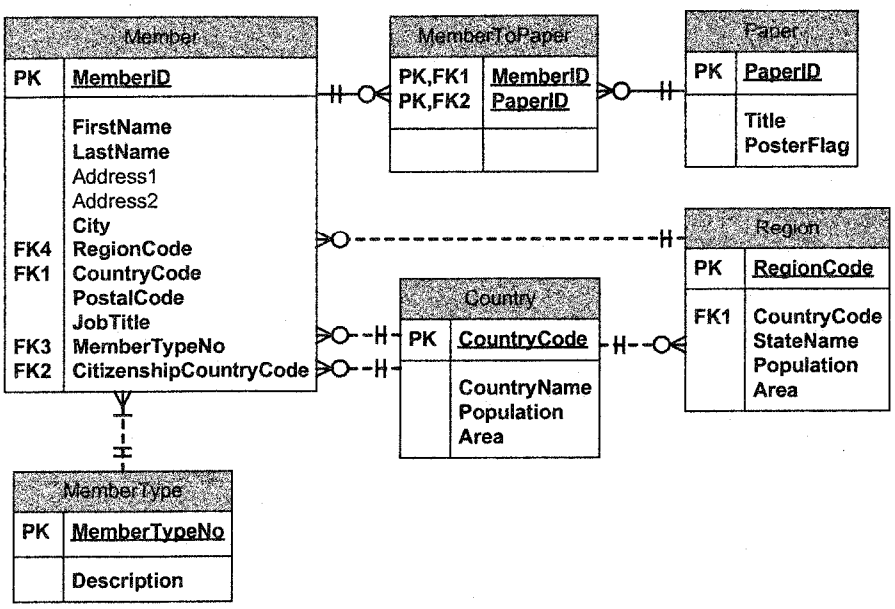

Figure 1 - An entity relationship diagram for the WSC spring meeting. Primary keys (PK) and foreign keys (FK) are designated for each entity.

Member : Table

MemberID	FirstName	LastName	City	RegionCode	CountryCode	PostalCode	MemberTypeNo	CitizenshipCountryCode
6	Kelly	Elder	Fort Collins	CO	USA	80526	0	USA
7	Thomas	Painter	Boulder	CO	USA	80304	0	USA
(AutoNumber)								

Record: 3 of 3

Figure 2 - Table and sample data for WSC member information.

MemberType : Table

MemberTypeNo	Description
0	Regular Member
1	Lifetime Member
2	Student Member

Record: 4

Figure 3 - Table and sample data for types of WSC members.

Country : Table

CountryCode	Name	Population	Area
CAN	Canada	31902268	9976140
USA	United States of America	280562489	9629091

Record: 1 of 2

Figure 4 - Table and sample data for country information. The area of the country is given in square kilometers.

Region : Table

RegionCode	CountryCode	StateName	Population	Area
AB	CAN	Alberta	2974800	661190
CA	USA	California	33871600	403968
CO	USA	Colorado	4056000	268656
*				

Record: 1 of 3

Figure 5 - Table and sample data for regions within countries.

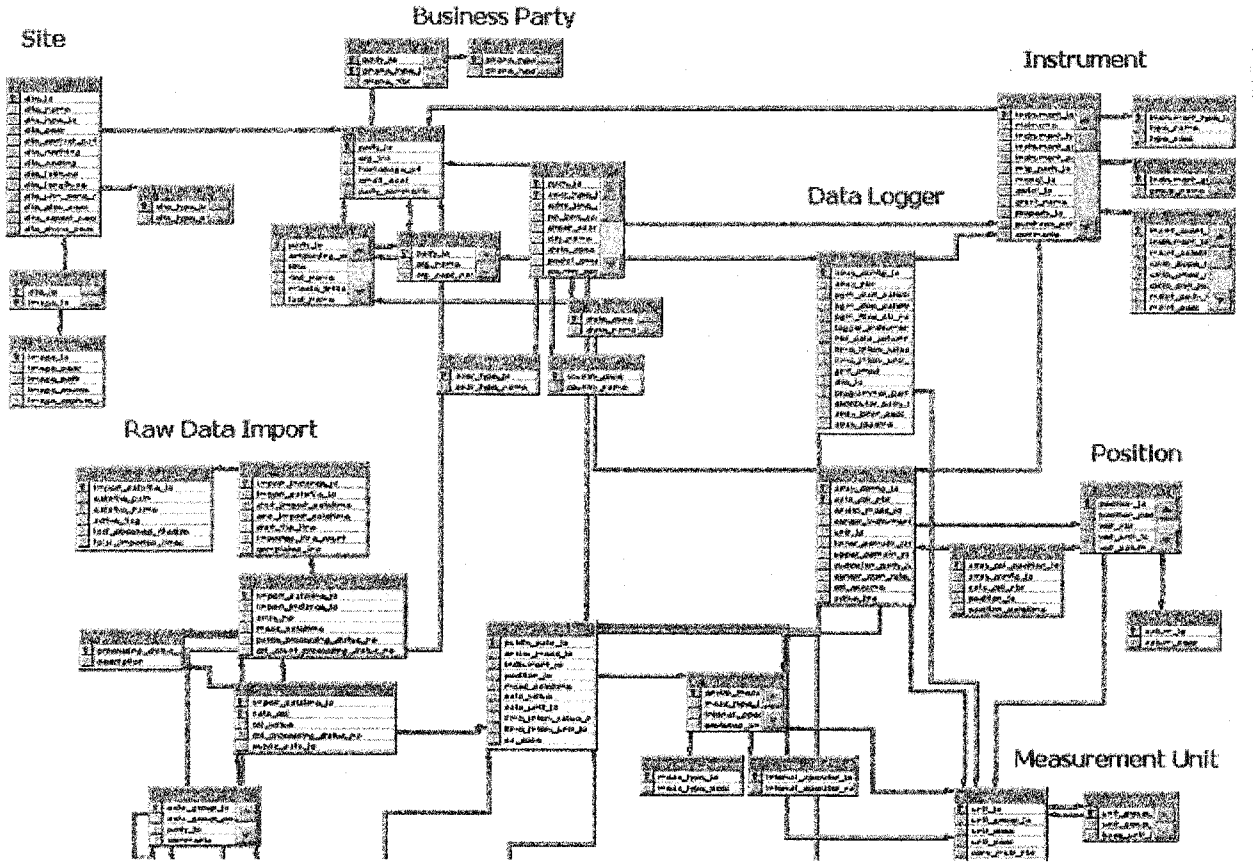

Figure 6 - Example of table structure and relationships for a large database. This example shows a partial list of table used by the Mammoth Mountain Energy Balance Monitoring Site database.

RELATIONAL DATABASE MANAGEMENT SYSTEMS

Databases are designed to efficiently store data, but additional tools are often needed for adding data to or extracting data from the database. This set of tools, including the database, is referred to as a relational database management system or RDBMS. A diagram of an example RDBMS is shown in Figure 7.

Several methods can be used to import data into a database. For small data sets, data can be typed in directly to the database. For larger datasets, data can be imported from a text file or spreadsheet.

Similarly, a variety of methods can be used to extract data from a database. The flexibility of data extraction is one of the most significant strengths of storing data in a database. Data can be exported into text files, passed to programs written in any programming language, and/or directed to a web page.

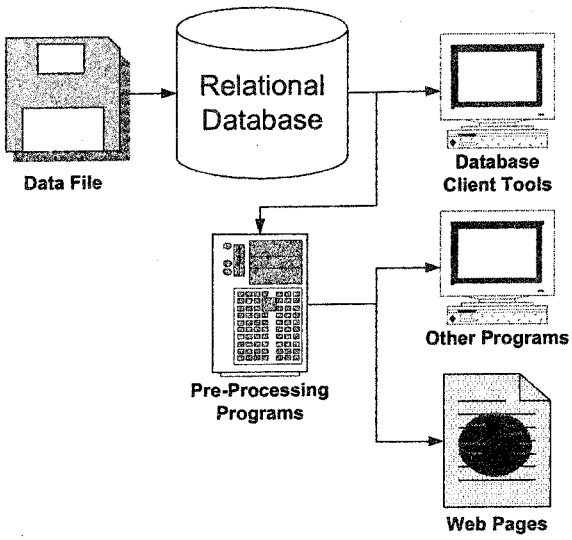

Figure 7 - Example database system diagram.

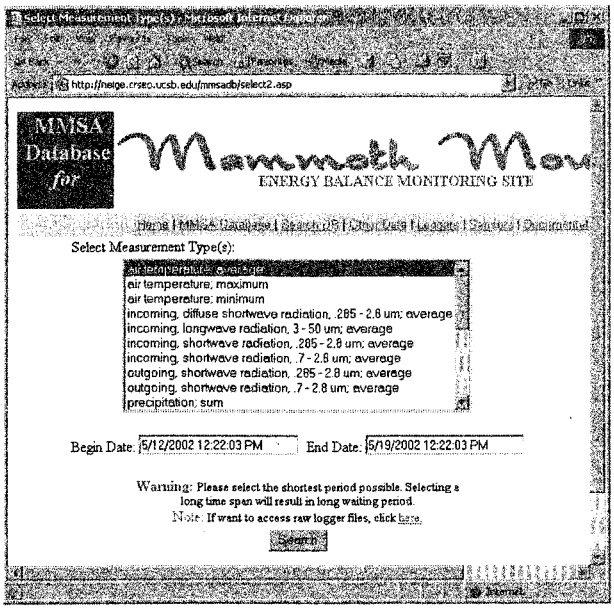

Figure 8 - Example of a dynamic web page that incorporates data stored in a database

Web pages are likely the most popular ways of distributing database information to a wide audience. Distribution by web pages potentially allows anyone with a browser to access the data. Figure 8 shows an example of a web page that displays data stored in a database that allows the user to view and download environmental data from the Mammoth Mountain Energy Balance Site database. This web pages can be found at http://neige.bren.ucsb.edu/mmsa/.

Other examples of web pages that access environmental data are:

- Snotel Data (http://www.wrcc.dri.edu/snotel.html) [3],
- EPA Water Quality Data (http://www.epa.gov/storet/) [1],
- USGS Surface Water Data (http://waterdata.usgs.gov/nwis/sw) [4].

SUMMARY AND CONCLUSIONS

Relational databases are an effective tool for storing and managing large data sets, especially when the data have a constant format. Data stored in a database can be distributed to a wide group of users, either through web pages or by means of a direct connection to the database. Because databases can reside on a single server that can be accessed by many users simultaneously, they reduce or eliminate the need for making copies of data sources.

Relative to other methods of data storage, databases require more time and knowledge for the initial setup, due to the need for data modeling and database table creation. For simple datasets this prior setup can make database systems more costly and slower to implement when compared to other storage methods (text files, spreadsheets, etc.). For anything more complex, the requirement of prior planning may have benefits in directing the collection of the data to be stored.

ACKNOWLEDGMENTS

Many of the examples given in this paper were created using funding from NASA-EOS.

LITERATURE CITED

1. Environmental Protection Agency. EPA Storet System Homepage. [web page]; http://www.epa.gov/storet/. [Accessed 15 May 2002].
2. MMSA Database for the Mammoth Mountain Energy Balance Site. [web page]; http://neige.crseo.ucsb.edu/mmsadb/. [Accessed 15 May 2002].
3. Western Regional Climate Center. Snotel Data. [web page]; http://www.wrcc.dri.edu/snotel.html. [Accessed 15 May 2002].

4. United States Geological Survey. USGS Surface Water Data for the Nation. [web page]; http://waterdata.usgs.gov/nwis/sw. [Accessed 15 May 2002].
5. The University of Texas at Austin. Data Modeling. [web page]; http://www.utexas.edu/cc/database/datamodeling/index.html [Accessed 25 July 2002]