Elizabeth Shay. Review and Refinement of a Database for Environmental Management Systems. A Master's paper for the M.S. in L.S. degree. August, 2000. 36 pages. Advisor: Stephanie W. Haas.

This paper describes a research project that employs an Microsoft Access database to track the environmental management systems (EMS) at participating industrial and municipal facilities. Two existing databases, which record the pre-EMS state of a facility and the EMS plan each facility proposes, were reviewed critically. A prototype database demonstrates the integration of the two existing databases into a single database, to improve public access and promote data consistency. Recommendations to the project leaders relate to database design and data management.

Headings:

Database design

Database management

Web databases

Environmental Management Systems

Environmental databases

REVIEW AND REFINEMENT OF A DATABASE FOR ENVIRONMENTAL MANAGEMENT SYSTEMS

by
Elizabeth Shay

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science

Chapel Hill, North Carolina

August 2000

Approved by:

_____

Advisor

TABLE OF CONTENTS

Introduction

This project involved critical review of two existing databases, integration of these databases into a single database, testing of the new database to ensure robustness, and documentation of the entire process to aid staff in using the final product. Because the project involves researchers from three different entities, and because most of the database design and management was executed by the database administrator in Washington, D.C., this project was primarily a behind-the-scenes effort to refine an extant system rather then physically implement the prototype. Experimentation on various aspects of the database design was reported to the database administrator in Washington and the key project leaders at the University of North Carolina in Chapel Hill, along with recommendations for design and implementation. This paper summarizes the project's history and goals, describes the design constraints presented by personnel and equipment, and provides examples of the kinds of information collected by the researchers. An explanation of construction of the integrated database is followed by recommendations for changes to the structure of the databases and strategies for managing the databases in the future.

The National Database on Environmental Management Systems (NDEMS) is a joint effort of the U.S. Environmental Protection Agency (EPA), lawyers and researchers at the Environmental Law Institute (ELI) in Washington, D.C., and faculty and doctoral students in the Curriculum in Public Policy Analysis (PUPA) at the University of North

Carolina at Chapel Hill (UNC-CH). It seeks to describe and track environmental management systems (EMS) implemented by facilities participating voluntarily in the program. Because the project receives federal funding, the information must be made freely available to the public; at the same time, anonymity has been guaranteed to all participating facilities.

Initial design of the project began in 1997; database construction, data collection and entry, and quality control began in 1998 and continues today. The first set of data was released to the public in June, 2000, with the second set slated for public release in early 2001. As the project evolves, new components may be conceived, designed, and launched. Eventually, the research team hopes to make the database interactive, allowing direct data submission and extensive data mining.

The researchers originally conceived of the project as a large database with three related but separate sets of data: a baseline component to describe the participating facilities through demographic and other descriptive information about the facility and its surrounding community; an EMS design component containing detailed information on the EMS the facilities design and implement; and finally an update component to track changes in the facilities' operation and performance following EMS implementation.

However, as the project developed, the baseline and EMS design data were entered into separate databases, with the location of data for future updates still undecided. The two databases have some features in common: each relates to a protocol that was mailed (physically or electronically) to participants, who returned the data to

the Chapel Hill staff for manual entry into the databases, and each ties all the tables together through the Facility ID number, which appears in a pictorial representation of the related tables as a “vanishing point” behind the screen because of anonymity provisions (Appendix E).

Because the baseline database was released to the public during this review process, it made sense to explore changes to database structure and design in the background, to allow a seamless transition to an integrated database in the future.

By spring, 2000, the baseline database was complete and full of data. The researchers found this database to be adequate for the tables, forms, and reports they required, although they made numerous alterations to the baseline database in order to accommodate unexpected data and to execute queries successfully.

While the baseline database was being prepared for public release in spring/summer 2000, the researchers tested the EMS design database and began to enter data. Problems in database design that came to light in the course of using the first (baseline) database helped the staff make design decisions that produced a better structure for the second (EMS design) database.

The project manager determined that a single integrated database containing all information is still the preferred format to serve public users who want to download the data from the project web site (http://www.eli.org). The project manager in Chapel Hill decided to use the sound structure from the second (EMS design) database as the parent database, incorporating all the data from the first (baseline) database, and allowing room

for update information. This project, which satisfies in part requirements for the master's program at the School of Information and Library Science (SILS) at UNC-CH, reviewed the two existing databases, constructed a prototype of a new, integrated database, stress-tested the integrated database with a set of deviant data, and offered recommendations for current and future team members who will use the new integrated database.

National Database on Environmental Management Systems

Research goals

The researchers who developed and maintain the National Database on Environmental Management Systems (NDEMS) are using database technology to track the experiences of industrial facilities that adopt environmental management systems (EMS). Their project, with funding from the U.S. Environmental Protection Agency (EPA), seeks to evaluate the impact of EMS systems implemented by the voluntary participants on the environmental and economic performance of these organizations.

The 1996 publication by the International Organization for Standardization (ISO) of the international standards for environmental management systems known as ISO 14001 provided an incentive for more businesses to explore and implement EMS plans. An EMS is a formal blueprint that describes how an organization can identify and reduce its impact on human health and the environment. These plans offer businesses the chance to demonstrate a commitment to management practices that reduce social costs and consumption of resources, and which often have positive concomitant effects on the

facility's operating costs.

Following a procedure prescribed in the ISO 14001 guidelines, an organization may win certification as ISO-compliant by a registrar—a third party that itself undergoes training, then offers this service for a fee. Certification as ISO 14001-compliant may improve an organization's standing with customers, suppliers, regulatory agencies, and the general public. Some firms are finding that status as ISO 14001-compliant has a significant impact on consumers' opinions and economic behavior, as manifest in signs proclaiming this achievement mounted by some industrial, municipal, or other facilities for public display. Some markets, where public and government pressure for certification alters the business landscape, may see an increase in the number of businesses adopting EMS.

A well-designed EMS plan should support an organization's development as it continually revises its operations to minimize environmental effects. There is disagreement among regulators and environmental groups over how effective an EMS may be in directing an organization's efforts to reduce its environmental impacts, and whether an EMS can ever substitute, in full or in part, for regulatory oversight of a facility. Aware of the need for research on implementation of ISO 14001 EMS plans and their effects on the facilities, a working group with representatives from 10 states and the EPA established a pilot program to gather data from businesses that were considering implementing ISO 14001-compliant EMS for eventual certification, and who were willing to share data on the process and results. Participants include manufacturing and

agricultural organizations, as well as municipal and military facilities.

The staff developed standardized protocols for data collection and entry, in order to produce a data set of dozens of comparable case studies. The UNC-CH and ELI teams designed the databases and corresponding protocols for data collection, and trained staff at facilities and state agencies in using the protocols. The program enrolled around 75 facilities in 1997. By May 2000 they had received complete demographic and descriptive data for 51 facilities, and partial data for 13 more facilities that are likely to continue with the pilot program. In addition, 17 facilities that had submitted data and implemented an EMS system, but are not part of a state program, will serve as controls, to assess the impact of the state's involvement. With cooperation across geographical regions, bureaucratic levels, and economic activities, this project has the potential to collect a vast volume of data that may be of interest to a broad range of users, including researchers, citizens' groups, and government agencies.

Personnel and equipment

The EPA supports NDEMS with funding and a project officer, who devotes approximately half his time to this effort. In addition, each of the ten participating states has a state manager, whose duties take up 10-15% of a normal work week.

In Chapel Hill, two doctoral-level research associates devote 15-20 hours per week to the project, while two faculty members are committed at the 20% level. The team receives 10 hours of administrative support each week, and employs two undergraduate

assistants for data entry, each at 10 hours per week. Three laptops are available to the team, along with the PUPA curriculum's computer lab, where machines running Windows NT are loaded with Microsoft Access 97.

The Washington staff includes three research associates and one senior legal professional, each devoting somewhat less than half their time to NDEMS. They have a similar computer arrangement, with access to several laptops as well as institute computer facilities, and they use the same Access 97 software.

All changes to database structure on the "frontline," i.e., in the database available to the public for downloading, are made in Washington by the database administrator (DBA). At this time, the database does not accept incoming data from facilities (let alone from users, including potential mischief-makers), or support any other transactions aside from a download interface. Because it is not an interactive database, concurrency control is not a problem (Levene & Loizou, 1999, chap. 8). Currently, the periodic updates of structural alterations from Washington and date entry from the Chapel Hill staff are done through e-mail attachments. The ELI and PUPA staff do not have a common server or mutual access privileges to each other's servers. This dictates an unwieldy process for updating database structure and content, requiring frequent and detailed communication to ensure that everyone is using the same version. The team is discussing future strategies for handling updates, particularly with the database's increased size in structure (new update tables) and content (new data entered).

Information Collected for NDEMS

Baseline data

The NDEMS project eventually will collect three sets of data submitted by the participating facilities, which roughly describe the "before," "during," and "after" stages of EMS design and implementation. In 1998-99, the staff collected baseline information, which covered the three years before a facility implemented an EMS. The researchers sent the facilities paper or electronic copies (depending on facility preference) of the protocol for these baseline data, which describe the pre-EMS (or "before") operation of a facility, including such data as facility size and type, work force, foreign production or marketing, and demographic information about the surrounding community and landscape. The facilities return their data either electronically or as a paper form. Data entry may involve copying and pasting between electronic protocol and NDEMS tables, or manual keying from a returned paper form or print-out of an electronically submitted form, depending on the preferences of the data enterer. The UNC-CH and ELI staff spent much of 1999 checking the baseline data for accuracy and consistency.

The team elected to use Access 97 as its database management system (DBMS), because of the widespread availability of this software and its relative ease of use. The incoming data were entered into a Microsoft Access database with 15 tables, each keyed to a unique ID number assigned to each facility (Facility ID). Each table has a corresponding form, through which the UNC researchers input most of the data for each facility (a small portion of the data was entered directly into the tables, when the format

or content dictated).

As the baseline data are submitted, the NDEMS staff perform quality assurance and quality control (QA/QC) on the data by reviewing the completed protocols, and consulting with the facility contact person when necessary. Rather than leave fields empty, the data researchers use placeholders to indicate missing data, then send a memo to the facility requesting clarification or additional information. If the facility does not provide the missing information, the researcher enters “no response” or “facility did not respond” into the field. In addition, the researchers look for signs of problems in the content, such as illogical or improbable responses, or identifying characteristics that might compromise the facility’s anonymity.

When QA/QC is complete, the team sends the facility a print-out of its completed baseline data for review and final sign-off. If the facility does not respond to NDEMS within a specified time, the data are assumed to be complete. These data for 51 responding facilities make up the baseline database, which the NDEMS researchers completed and released to the public in June, 2000.

EMS design data

The second set of data, in which the facility describes its EMS design by responding to 15 sets of questions, covers the “during” stage of EMS implementation. The EMS design protocol, nearly 40 pages long (including nearly 10 pages of introduction, terminology, and other supporting information), went out to facilities in

August, 1999. Those facilities that had already begun completing earlier versions of the EMS design protocol before the final version needed a patch that folded their existing data into the final version of the EMS design protocol.

The questions in this protocol cover such topics as the rationale for adopting an EMS, objectives and targets, supply chain relationships, and monitoring and measuring of regulatory compliance. Many of the questions allow multiple responses, or extensive explanations entered in memo fields.

These data are entered into the 23 table/form pairs of the second (EMS design) database as they arrive, which may precede final sign-off on the baseline data. Data entry serves as a first step in quality control for EMS design data; when missing, incomplete, or vague answers appear, the NDEMS staff can contact the facility for clarification. Data entry and editing follows a procedure similar to that used with baseline data, with one or more rounds of communication between the researchers and facility before final sign-off on data. The NDEMS team hopes to have complete EMS design data from all 51 facilities by December 31, 2000.

Update data

The third set of data collected for each facility illustrates the "after" conditions in an organization that has implemented an EMS. The EMS design data for any given facility must be complete *before* the update protocols are sent out, to prevent distortion in reporting of the EMS design data. Eventually, the project will include an update for

each table/form in the baseline component, to be submitted twice each year.

The data collected for this component includes such information as changes in regulatory status, emissions levels, or environmental indicators. These data should show how a facility's environmental and economic performance changed after it implemented its EMS. Since the best EMS designs will be flexible and guide an organization through many changes, the update protocol should allow for multiple updates (not just the most recent) to track changes over time. One update protocol (corresponding to the first table in the baseline component) is complete, while work continues on the upcoming protocols.

Early Database Construction and Review

The research team first conceived of the project as a database with three related components, each covering a different stage of the research. In practice, the first two components were created as separate databases. The databases were built through collaboration of the PUPA team in Chapel Hill and the ELI staff in Washington, who were heavily involved in writing queries to run from the baseline database. Because data were entered as the facilities submitted their baseline protocols, the database already contained a significant volume of data when snags with the structure began to appear.

While enrolled as a student in a database class at UNC-CH's School of Information and Library Science, I volunteered to work on this project, and was assigned first to critique the baseline database. A February 2000 memorandum to the UNC team (Appendix B) summarizes my comments and suggestions on the structure of the first

database (later, shorter communications covered other lingering snags and new observations). Many of my suggestions were relatively minor details that related to consistency in nomenclature. However, I did point out several substantive problems with implications for future use of the database, and discussed my concerns with several staff members. These problems included spaces in field names, problems with field type (text vs. number or date) and field length, inconsistent data types between tables, and absence of required fields. Because the database already contained complete records for dozens of facilities, and because many dozens of queries already had been run on the database, the research team made only those changes that would not negate the work already completed on the tables and queries.

When the research team completed the second (EMS design) database and prepared to enter data, I reviewed this database and again submitted comments to the UNC-CH team. These comments (Appendix C) related largely to organization of the tables and to labels for tables and fields, and in some cases to data types of fields (memos and text fields, yes/no boxes, and others). This database had a sounder structure than the first (baseline) database, and contained very little data. The DBA in Washington made substantial and continuing improvements to this database, in response to my communications as well as his own experience in working with the baseline data.

While tinkering with the structure of the databases and refining the protocols for the third set of data (updates) to be collected, the research team decided to return to the original plan of creating one integrated database to hold all the data, from the baseline

demographic through the EMS design data, with new tables for the update information. This provides the public with the convenience of a single database, from which it is easy to run queries and reports. Moreover, an integrated database will help enforce data discipline, so that data covering common material is handled consistently. Because the team was generally satisfied with the structure of the EMS Design Database and the appearance of the forms, they elected to use that database as a starting point, move into that database all the tables and forms from the baseline database, and build in room for the information that would have gone into an update database, had it been created as a separate entity.

Construction of an Integrated Database

Preparation

To prepare for building a single, integrated database from two existing databases (and one available protocol for future update information), I first reviewed the current baseline and EMS design databases, to see which suggested changes had been made. Then I discussed with the UNC-CH researchers the EMS design database, taking careful notes on what they consider the most critical and useful features of that database. We also looked through the baseline database to flag any fields that should be altered or rearranged before being inserted into the new database. While reviewing these databases, we discussed the user interface and possible refinements to it.

We concluded that, because the baseline database already had been released to the

public, the baseline data should be combined with the EMS data with as few changes as possible. The database administrator in Washington created a user interface on top of the baseline tables and forms to direct users to their desired information. I created a similar interface for the EMS design component, and changed the backgrounds on the user interfaces to match the backgrounds of their respective forms (yellow for baseline, blue for EMS design). In making these adjustments, I referred to Jasco's (1999, chap. 1) discussion on principles of interface design, which stressed the importance of intuitive layout, since most users do not want to consult documentation, even when available online. Aside from suggestions I submitted relating to minor issues of content and layout (misspellings, inconsistencies in terminology), no other major changes are planned for the current presentation.

In preparation for merging the two databases, I made a copy of the current EMS design database to use as my working template of the new integrated database. Because the tables and forms in each of the two databases (baseline and EMS design) were numbered sequentially, an integrated database would require some device to distinguish the tables and forms of the two different major components. My solution was to give all baseline tables and forms a prefix of "BL," and those of the EMS design data the prefix "EMS." Starting with the EMS tables and forms in the current EMS design database, and working with pairs of tables and forms, each with the same original number, I first changed the record source in the form's properties list, then changed the form name and corresponding table names.

In this procedure, I opened the first EMS design form ("Rationale for adopting an EMS") in design view, went to the properties list under the View menu, changed the name of the record source by adding the "EMS" prefix, and saved the form. Then I renamed the form with the prefix "EMS," and renamed the corresponding table with the same prefix. The newly named form now called up the corresponding newly named table. Initially, I did a single form/table pair at a time. Once I was comfortable with the procedure, I changed all the record sources in the forms' properties lists and renamed all forms in the same manner, then renamed the corresponding tables. After completing this procedure, I scrolled through the forms to verify that each form called up the correct table with the corresponding name.

On my first attempt at this procedure, I encountered problems with several long form/table names. If the record source in the form properties list accepts the new, longer name and the changes are saved, but the addition of the prefix makes the name too long for the table to accept, the form cannot call up the table, nor can it be opened for editing. In these cases, I was forced to bring in a new copy of the form and table from outside the database. With that knowledge, I watched for long names that might cause this problem, and shortened the name *before* editing the table and form names.

Adding baseline data

After I renamed all EMS design data tables and forms with the prefix "EMS," I imported the baseline data into the EMS design database. With the EMS design database

open, I chose “get external data” under the File menu. Then I opened the most recent copy of the baseline database, selected “all tables” and chose the options “Import relationships” and “Import definitions and data.” Clicking OK copied into the EMS design database all the baseline tables, with the relationships intact. Like the EMS design database, the baseline database ties all tables together at the Facility ID number, which is kept out of the database to ensure anonymity, and thus shows up as a “vanishing point” to which all tables point. I repeated this import procedure for baseline forms and queries.

After insertion of the baseline data, the relationship window in the EMS design database shows two separate sets of tables, one each for EMS design and baseline data, each tied together at the Facility ID. The first update protocol, produced from the facilities demographics information, floats alone, not yet tied to either set of data (Appendix E).

After importing the baseline data, I went through the same process of renaming forms and tables, giving each a “BL” prefix. As before, I took care to change record source and form names first, and to watch for long names, to avoid having an unduly long new form/table name force me into an endless loop of being unable to call up a form without changing the corresponding table name, which can’t be done until the record source and table name match. Again, I scrolled through all the forms after renaming forms and tables, to verify that each newly named “BL” form calls up the appropriate newly named table.

With all the tables and forms in a single database structure, and each form pulling up the correct table, I turned to the user interfaces for the two components. In the user

interface created by the Washington DBA for the baseline data, I first renamed all the control buttons to match the newly renamed forms, and went into the code for each form to change the name of the table each form calls up. I changed the background of the baseline user interface to yellow to match the baseline forms. Then I followed the same pattern to create a user interface for the EMS design forms, and gave it a blue background to match the EMS design forms. I made the form "read-only" by clicking "No" in the properties list after the options "Allow edits," "Allow additions," and "Allow deletions," to prevent alteration of records by users.

Update data

Using a paper copy of the update protocol for the Facility Demographics (Table BL0 from the baseline database), which requests information about changes in facilities demographics and operations *after* EMS implementation, I created tables to report updated information. The updates are submitted on a specified schedule, twice a year. Because the NDEMS wants to track changes for an extended period with multiple updates, I designed a table in which the Facility ID and date of a single update comprise the primary key, and thus uniquely identify each update, assuming no more than one update in any given day—a safe assumption for updates that are scheduled twice yearly. The remaining fields of this first update table record changes to such parameters as foreign production and marketing, and relationship of the facility to a larger entity.

Certain fields in the first update table would have been good candidates for all-key

tables to reduce the number of null fields. For example, I considered creating new all-key tables (Facility ID and country) to tie facilities to multiple foreign countries where they have production or marketing, since few facilities will use all available fields allotted these attributes. However, because this would depart from the structure of the baseline database, which contains fields for up to ten countries for each of these attributes, and would require undue complication to reconcile these two tables, I elected to leave the update table untouched, identical to the baseline table to which it relates.

The NDEMS team expects to create an update protocol for each table in the baseline component; in addition, some of the EMS design tables may be updated. As these protocols are drafted, tables can be created for each by copying the parent table, naming them after the parent table with a "UD" prefix, and adding, deleting, or altering fields to match the questions in the update protocol. Copying and editing the parent tables is easier than simply building them anew. Moreover, this procedure will make it easier to keep data types, field lengths, input masks, and other details consistent between initial and update data, thus improving the quality of the data.

Testing the integrated database

To test the new, integrated database for robustness, I performed several trials designed to verify that the data import/export procedure went smoothly, and make certain the database does what the researchers want.

After the Washington DBA creates or edits any table or form, he sends it to the

researchers and their student assistants, who identify problems quickly by entering the large volumes of data common in the protocols. They send feedback to the DBA, who adjusts the tables and forms as necessary. Because the data entry currently is done manually and supervised by the researchers who designed the protocols and will be analyzing the data, the tables and forms are relatively free of snags by the time of public release.

To test the soundness of the integrated database, I went through each table and form, one at a time. I reviewed each table in design view, looking at the various data types and field lengths. Then I returned to the table view, and challenged the database structure by entering illogical or impermissible data. For example, I entered text and currency into number and date/time fields. I also attempted to enter into date fields illogical data such as more than 31 days, more than 12 months, and years of 9999. I tried to create records without required fields (usually the facility ID, which serves as the primary key for each table in this database). Finally, I entered more characters than the field length allows, and attempted to copy in large volumes of data into text fields.

The database generally performed well under these tests. Most entries of inappropriate data were denied; some exceptions, stemming from problems mentioned in the original critical reviews of the database, are described in Appendix D. The database prevented the creation of dummy data by requiring that new records in most tables refer to a record already existing in the major baseline or EMS design tables, which serve as central tables for the two components.

Nearly all data entry proceeds through the forms that correspond to the underlying tables. These forms often offer a choice of answers for a text field, with the field length set at the length of the longest possible answer. For example, some questions have a choice of answers of “some difference,” “no difference,” or “no response,” with a field length of 15 characters. The table allows entry of any 15 characters, while the form allows only choices from the menu of answers. Since most data are entered through these forms, there is less chance of insertion of irrelevant data into these fields. The forms are now inactive; i.e., when the user downloads the database, s/he can see all the questions, menu choices for answers, and data, but cannot add, delete, or alter records.

Design Considerations

Several chronic design problems stem from non-technical issues with technical implications. The question of size involves a trade-off between completeness of data and ease of database downloading. At the same time, the complexity of the evolving database sets up a tension between service to sophisticated users who would value the potential analytical power of running their own queries and reports, and to less skilled database users who may desire canned queries and reports with commonly sought information.

In addition, the mandate for public access to data issuing from federally funded research raises concerns about integrity of the “official” version of a database to which several groups of people have access, and about control over data once it is released to the public.

Database size

The size of the database continues to be a concern, because of the tension between the contradictory goals of keeping the data in a single, integrated (and therefore large) database, on one hand, and keeping the size of the database manageable for easy downloading, on the other. These two competing goals both serve the project's overarching philosophy of public access. Because the researchers are committed to collecting all information the facilities are willing to provide, the protocols have ample room for extensive (and in the case of memo fields, nearly unlimited) text with which facilities may report, describe and explain their data. At the same time, this commitment to quantity, which often supports quality (completeness) of data, makes the database unwieldy and awkward to manipulate. Jasco (1999, chap. 5) describes other situations where field length is a concern.

The 20 MB of space the integrated database occupied at the time of this writing included completed baseline data for all 51 facilities that had signed off, partial EMS design data for some facilities, and one (empty) update table. Still to come is a large volume of data to complete the EMS design component, and many update tables (structure and content) for protocols not yet written. Clearly the issue of size looms large for this project, and begs an evolving strategy to balance comprehensiveness of data with access and ease of use for the public. Occasional data transfer problems in the course of building the prototype (i.e., importing the baseline data into the EMS design database) may have been related to the size of the database, and suggest that the general public may

find it difficult to work with a single large database.

The current public release of the baseline data handled from the ELI web site as a full database download has the potential to deter users who are not facile with Access or similar software, thus limiting the usefulness of the data for some people. The current all-or-nothing download procedure may give way eventually to an interactive web-based database allowing selective downloading of desired information only.

Because the baseline database was released to the public in June, 2000, with the EMS design database slated for release sometime in 2001, the researchers may wish to experiment with the prototype integrated database in the background before deciding to give up the separate databases in favor of a single entity. If the integrated option is implemented, presenting the integrated database to the public at the time of EMS design release may be the most elegant scheduling option.

Integrity of database

The researchers continue to discuss the problem of "locking" the database, to prevent tampering with the data. In fact, there is little that can be done to prevent manipulation of the data once a user has downloaded it from the ELI web site (Nance, 1998). Indeed, security is a major concern among most organizations contemplating or currently offering database access via the Internet (Malong, 1997). The master database itself can be protected from tampering by a variety of strategies, in which the ELI staff are well versed. They also have taken steps to make downloaded data resistent to

unintentional alteration. For example, only “select queries” are possible. Moreover, the opening interfaces are read-only forms, which take the user to the various forms rather than tables, and display all questions, response menus, and data, but do not permit additions, deletions, or alterations of records.

Of course, sophisticated and determined users could do more to manipulate the underlying data. Because the researchers cannot prevent users from changing data after downloading, the team may want to post a disclaimer on the web site, perhaps requiring an acknowledgment click by users, stating that integrity of the data is not assured once the database has been downloaded. This may be useful in cases of inadvertent alteration of data or intentional or malicious misrepresentation or manipulation of data, both of which potentially could be used to produce spurious findings using government-sponsored data.

Currently, the only control on access to the database is an interface that collects user data before allowing downloading. The download form inserts a cookie into the user’s computer that identifies the computer by IP address. The required fields in the download interface include name, title, affiliation, and an explanation of the user’s interest in the data. While a malicious troublemaker could falsify such data, this database is unlikely to attract serious hackers, and the self-reported data will offer the research team some insight into who is using the data, and for what purpose.

The NDEMS made a conscious decision not to write and post a user’s guide to the database, in part because the budget does not provide resources adequate to cover this

effort, and with many staff already devoting more than their allotted time to the project, they do not want this to become an expected part of their work. Moreover, they are not a service organization, but rather a research entity dedicated to collecting and analyzing data; as such they are focused on the scholarly aspects of the study and the goals of their legal consultants, faculty researchers, and doctoral students.

As the database grows in size and variety of content, it is tempting to label it a "data warehouse" in keeping with the description by Garcia-Molina et al. (2000). In this sense, the Washington "frontline" database stores data from more than one source and allows access to users without permitting updates from sources outside the small distributed system; at the same time, it accepts updates from the Chapel Hill site, which are processed later for less frequent updating to the public interface. Elmasri and Navathe (2000) have a more elaborate and stringent definition of a data warehouse, which exceeds the scale and functionality of this project.

Peckham (1999) cites sound design technique as a critical element in managing databases during this time of rapid change in standards, which recently have evolved toward web functionality largely without the participation of database designers. Fryer (1998) also stresses the need for good early database design to allow a system to grow. In his description, scalability of databases into data warehouses requires consideration for the volume of data and the number of concurrent users, as well as the environment and the system's ability to support the desired functions. Whether or not NDEMS qualifies as a data warehouse, the philosophy of designing for future growth and new demands is valid.

Recommendations to NDEMS

The following recommendations, which stem from my experience with the prototype and reading in the literature, may aid the NDEMS team in current and future management of this database:

- Before deciding on whether to integrate the databases into a single entity, carefully compare the two options. Examine the prototype integrated database in light of the goals of the project. Survey early users of the baseline and EMS design databases, to determine whether they can manipulate the data when downloaded as two separate databases, and what kinds of problems they have encountered with queries and reports. Have the data entry assistants enter data for several forms, and run queries and generate reports from both formats. Ask friends or colleagues working at home and offices with various kinds of hardware to try downloading each format.

- If the integrated database is chosen as preferable, follow the procedure described here to integrate the baseline data into the EMS design database (or vice versa). First, while the databases are separate, rename all EMS forms and tables (e.g., with an "EMS" prefix) and scroll through all forms to verify that they call up the appropriate table. Then import the baseline data into the EMS design database, and rename all the baseline forms and tables in the same manner (e.g., with a "BL"

prefix), and verify that forms call up the correct tables. Retain this pattern of using prefixes for the different components (baseline vs. EMS design), and retain parallel naming of forms and tables that comprise pairs, wherever appropriate.

- Create a user interface for the major components (EMS design and updates) similar to the recently created interface that opens the baseline database. This organizes the sections, and takes users to forms rather than tables; it also provides a buffer between the user and the base tables, which may hinder accidental or intentional alteration of data, if that remains a goal of the project. Alternatively, the researchers may want to keep updates nested together with their corresponding tables, for easy comparison, rather than segregating the updates from the other components.

- When creating new tables for updates, use an identifying prefix (e.g., “UD”) to distinguish it from baseline and EMS design data. Copy parent tables when creating new update tables, to ensure consistency of data types and field properties; then add, delete or edit where necessary to match the update protocol distributed to the facilities.

- Create a switchboard as the opening interface, one level up from the interfaces for the three (or two) components, which directs users to the major sections of the

database. This screen could include an exit to the tables, if the researchers want users to have such direct access to the underlying data.

- Use different colors for the different components of the database. The baseline forms are already yellow; this color could be extended to the backgrounds for reports and the user interface. Rose or blue could represent the EMS design component, while green would be an appropriate symbolic choice for the updates, representing improved environmental performance.

- To warn users about the possibility of data manipulation after downloading, post a disclaimer on the web site, perhaps requiring a click by users, explaining that integrity of the data is not assured once the database has been downloaded.

Summary

This project offered an opportunity to put database design principles to work in a system with several distinctive non-technical features. While the database structure itself is relatively simple—a series of tables with the same primary key, each with one corresponding form, the goals of comprehensiveness and broad public access pose special challenges.

Because one component already was nearing completion and public release while subsequent components were still in the development stage, many design problems

required creative solutions that bridge the gap between old and new, and needed to accommodate structural features in the integrated database that would not be desirable in a newly designed database.

In a collaborative work environment such as NDEMS, with several geographically remote groups of researchers and no hard network connection, maintaining integrity of the database and control over data is difficult. Evolving projects that change hands periodically are particularly challenging.

Thoughtful database design promotes robust systems, which can absorb occasional bending of design principles in the service of the particular requirements of a given database. Databases, after all, were created by humans to solve problems and manage information, and as such must be viewed as servants rather than masters.

This project might have benefitted from the services of an experienced database designer in the early stages, although even this measure could not have prevented evolution of the project or anticipated all problems. The NDEMS team's excellent DBA, and the team's cooperative and flexible ethos, have made this project successful despite the many challenges.

References

Elmasri, R., & Navathe, S. (2000). *Fundamental of Database Systems*, 3rd ed. Menlo Park, CA: Addison-Wesley.

Fryer, R. (1998). Data-warehouse scalability: How well a database system grows over time will be determined by decisions made in the initial design. *Byte*, *23* (6), 63-64.

Garcia-Molina, H., Ullman, J. D., & Widom, J. (2000). *Database System Implementation*. Upper Saddle River, NJ: Prentice Hall.

Jasco, P., & Lancaster, F. W. (1999). *Build Your Own Database*. Chicago: American Library Association.

Levene, M., & Loizou, G. (1999). *A Guided Tour of Relational Databases and Beyond*. London: Springer.

Malong, L. G. (1997). *Database connectivity via the Internet: A study of current practices*. Master's paper, University of North Carolina-Chapel Hill.

Nance, B. (1998, October 15). Improving Data Access Security. *Network Computing*, *9* (19), 122-124.

Peckham, J. (1999). Data for the masses. *Journal of Database Management*, *10* (2), 35-37.

Appendix A: List of Acronyms

DBA: Database administrator

DBMS: Database management system

ELI: Environmental Law Institute

EMS: Environmental management system

EPA: United States Environmental Protection Agency

ISO: International Organization for Standardization

NDEMS: National Database on Environmental Management Systems

PUPA: Public Policy and Analysis curriculum

QA/QC: Quality assurance and quality control

SILS: School of Information and Library Science

UNC-CH: University of North Carolina at Chapel Hill

Appendix B: Comments on baseline database

In all tables, there are spaces between words in field names. [This] may have been done to facilitate creation of forms with nice-looking titles.The project directors need to know that the spaces are not a problem in Access, but might be if they ever move the data to a different database management system.

In all tables, the "Preparer ID" field is indexed. This seems a funny field to index

In all tables, the "Entry date" is a 50-character text field. If this field were a data/time field, they could do queries using dates, to exclude or include records from certain periods.

In all tables, the description for "Entry date field" has "data entry" spelled incorrectly.

Table 0 (Demographics)
The "Facility ID" field should match the same field in other tables. Table 1 describes this field as "Facility unique ID number" ("Unique facility ID number" reads better).

The "Ownership" and "SIC" fields have long fields of 50 characters . . .Could the field lengths be shortened?

There may be a mistake in the Prod Country and Mark Country fields. The 10 "Prod Country" fields all have a description saying "Foreign production in this country." The "Mark Country" fields say "Market product in this country" in the first three fields, but switch back to "Foreign production in this country" in "Mark Country" fields 4-10.

The "Employees" and "Population" fields are each 50-character text fields. If these were number fields, they could be used to sort by size of work force or population

Table 1: Management Systems Baseline
"External QA check done" field has a misspelling in the description (quality assessment").

The "Gap analysis" field say it's a yes/no box, but looks like a text box.

Table 2-table
The "Performance indicator number" field has "performance" misspelled in the field name. There is no description. Do the clients know what this number is?

Table 3-Compliance baseline
There is no primary key.

Appendix B, cont.

The "Facility ID" should match the same field in other tables (such as 0 and 1), and be a number rather than text field.

Table 3-report
The primary key is "Regulatory Status Date Index," an autonumber. Then the next field, not identified as primary key, has the same entry, but with a misspelled [and rearranged] field name ("Statu Regulatory Date Index") and a number data type. Should the second field be removed? Also, the field name of the primary key is so long, perhaps it could have been called RSDI, with a note in the description giving the full name.

Table 3—table 3
The "Reporting year" field is a 50-character text field. Why not a 4-digit number field?

Table 4
The "Facility ID" is not a required field. In fact, there is no required field in this table. Is this to protect the anonymity of the clients?

There are several fields that have small numbers as entries, but have 50-character text fields as the data type. They are "Involve suppliers," "Involve customers," "Reward," and "Training." Would it be useful to have these as number boxes? Also, "Accounting" looks like it should be a yes/no box.

Table 5
The "Facility ID" field is not required, nor is any other field in this table. Is this a security measure?

Table 6-reports
The "Facility ID" field is not required, nor is any other field in this table. Is this a security measure?

Appendix C: Comments on EMS database design

General comments:

The EMS Design database includes 15 tables and 15 corresponding forms,which are hard to navigate because of the numerous fields with long field names (column headings) that run across the top of the screen in table view. Since nearly all the tables run off the right side of the screen and require users to scroll to see all the fields (columns), it might be helpful to either shorten the field names or make the field name labels two or three lines deep so that field names take up less horizontal space. This may be particularly useful in the case of current Y/N data types, or in those that might be changed from text fields toY/N.

As in the baseline database, the field names have spaces in them, which Access allows (and which may be convenient for pasting into forms entries), but they may be problematic if these data are ever migrated to another database management system.

None of the tables or forms have descriptive titles that would guide users to tables or forms of interest. If tables and forms are given descriptive titles, pairs of tables and forms still should have related titles (as they do now, with Table 1 corresponding to Form 1).

In all the tables, the facility ID is a 50-character (default size) text field. If the field size is limited to 3, and no facility ID numbers exceed 3 characters, this can help prevent entry of data into this field meant for another field. The same is true of other text fields that should have short entries (e.g., preparer ID), even when they are left as text fields because no arithmetic operations will be performed on them.

None of the tables have a required field, although all use facility ID as the primary key. This may be acceptable, if you are interested in recording any data you can get, even it you can't directly tie it to a particular facility.

There are a mixture of Y/N and text data types used for questions in the tables. It appears that a 50-character text field usually corresponds to questions in the related forms that offer a yes/no menu box (sometimes appearing as “Y” and “N” instead of “yes” and “no,” which could be standardized for consistency), while the Y/N boxes that are used in tables correspond to click boxes in the related forms. If there is no reason for the difference in presentation, the text fields for questions in the tables that correspond to yes/no menus in forms could be changed to Y/N data types in the tables. This might prevent inappropriate data from being entered mistakenly in those fields.

Appendix C, cont.

Some of the questions used as field names sound awkward. For example, in many cases, there is no question mark to identify it clearly as a Y/N question, and if it is a text field, any kind of text could be entered there. In addition, in many cases the initial "Does" is dropped off the beginning of the sentence, resulting in a field name such as "facility have a procedure. . .," which is quite awkward. Perhaps fields names could be cleaned up to something like "Facility has procedure?"

Many of the titles have a series of these Y/N questions, often alternating with a "please explain" field. While the Y/N questions are too long (except those that are set up as Y/N boxes rather than text), the "please explain" fields are too short, with a 50-character text field. The forms that correspond with such table fields show a nice big box for "please explain" answers, but only allow 50 characters to be entered. These might make good memo fields, although memo fields open up the possibility of excessively long answers and the attendant problems with unwieldiness of the database and difficulty for users. The project staff members made it clear that they want to encourage long answers to such questions ("the more information the better") to capture the most comprehensive information possible; the technical database specialists suggested a longer text field instead of memo fields. This question remains open.

Some of the tables have fields for "Total person hours" and "Months" as text fields. If these were number fields, they could be used for comparisons that required computation.

Table-specific comments:

Table 1:
The field that begins "Shareholders' and owners' . . ." has a caption, unlike all the other fields.

The menu on the corresponding form offers "H," "M," "L," etc. as answers to some questions. But the table view for this table has "H" entries on the first line, and "No" entries on the second.

Appendix D: Comments on data testing

Testing the data for problems with data types and field properties turned up no major problems with data transferred into the integrated database.

The two components (baseline and EMS design) are each generally consistent within the given component's tables and forms, although they each have distinctive characteristics and problems.

The tables imported in from the original baseline database all have preparer ID defined as a number field (although no arithmetic operations will be performed on it). The data entry date is entered into a 50-character text field, which allows entry of various kinds of text without triggering a message requesting appropriate and logical information. In addition, several tables had other data-type problems.For example, in table BL0, the employee and population fields are meant to record the size of the work force and population of the surrounding community, but the 50-character text fields are too large and allow non-numeric information. The foreign marketing countries are memo fields, which is too liberal for country names (25 or 50 characters should suffice), and would allow accidental insertion of extraneous text. Likewise, Table BL1 has many text fields for recording duration of certain events that should be limited text fields rather than memo fields. Finally, Table BL3—Report 3 has a memo field for information relating to a date. Although it is too late to make this a date field (some information was recorded as “since . . .” or “starting in . . .”), a shorter text field would prevent insertion of extraneous text.

Despite these minor problems with data consistency, the overall structure of the database successfully prevents entry of dummy data by requiring new records in most tables to refer to a record already existing in the demographic baseline table.

The EMS design tables have a cleaner design, with all “Facility ID” fields defined as 5-character text fields and the “Preparer ID” field is a 2-character text field. These shorter fields may prevent some insertion of inappropriate data. Moreover, the entry date for new data is a date field with an input mask.

Because the facility ID cannot be null, dummy records cannot be inserted unless they refer to an already existing record in Table EMS1. The tables will allow text of any kind in many fields, while entry of data through the corresponding forms allow only certain choices. The field lengths of many text fields correspond to choices in the forms (e.g., no difference,” “no response”). On the rare occasions when the researchers enter data through the tables rather than forms, they must be careful to enter the categories specified in the forms rather than free text.The baseline forms are inactive, i.e., it is not possible to add, delete, or alter information. Users can scroll through the records and see the data

entered, as well as the menu choices for individual questions, but cannot change any data. The EMS design forms are still active; the team is working on locking these data.