U.S. Department of the Interior
U.S. Geological Survey

Prepared in cooperation with the
FEDERAL HIGHWAY ADMINISTRATION

Data Model and Relational Database Design for Highway Runoff Water-Quality Metadata

Open-File Report 00-480

A Contribution to the
NATIONAL HIGHWAY RUNOFF DATA AND METHODOLOGY SYNTHESIS

U.S. Department
of Transportation

U.S. Department of the Interior
U.S. Geological Survey

Data Model and Relational Database Design for Highway Runoff Water-Quality Metadata

By Gregory E. Granato and Steven Tessler

Open-File Report 00-480

A Contribution to the
NATIONAL HIGHWAY RUNOFF DATA AND METHODOLOGY SYNTHESIS

Prepared in cooperation with the
FEDERAL HIGHWAY ADMINISTRATION

Northborough, Massachusetts
2001

U.S. DEPARTMENT OF THE INTERIOR
GALE A. NORTON, Secretary

U.S. GEOLOGICAL SURVEY
Charles G. Groat, Director

For additional information write to:

Chief, Massachusetts-Rhode Island District
U.S. Geological Survey
Water Resources Division
10 Bearfoot Road
Northborough, MA 01532

Copies of this report can be purchased from:

U.S. Geological Survey
Branch of Information Services
Box 25286
Denver, CO 80225-0286

PREFACE

Knowledge of the characteristics of highway runoff (concentrations and loads of constituents and the physical and chemical processes which produce this runoff) is important for decision makers, planners, and highway engineers to assess and mitigate possible adverse impacts of highway runoff on the Nation's receiving waters. In October 1996, the Federal Highway Administration and the U.S. Geological Survey began the National Highway Runoff Data and Methodology Synthesis to provide a catalog of the pertinent information available; to define the necessary documentation to determine if data are valid (useful for intended purposes), current, and technically supportable; and to evaluate available sources in terms of current and foreseeable information needs. This paper is one contribution to the National Highway Runoff Data and Methodology Synthesis and is being made available as a U.S. Geological Survey Open-File Report pending its inclusion in a volume or series to be published by the Federal Highway Administration. More information about this project is available on the World Wide Web at http://ma.water.usgs.gov/fhwa/runwater.htm

Fred G. Bank
Team Leader
Office of Natural Environment
Federal Highway Administration

Patricia A. Cazenas, P.E., L.S.
Highway Engineer
Office of Natural Environment
Federal Highway Administration

Gregory E. Granato
Hydrologist
U.S. Geological Survey

CONTENTS

PLATES
(On CD in back pocket)

FIGURES:

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
LENGTH				
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
AREA				
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yards	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
VOLUME				
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
NOTE: Volumes greater than 1000 l shall be shown in m^3.				
MASS				
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
TEMPERATURE (exact)				
°F	Fahrenheit temperature	5(F-32)/9 or (F-32)/1.8	Celcius temperature	°C
ILLUMINATION				
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
FORCE and PRESSURE or STRESS				
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
LENGTH				
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
AREA				
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
VOLUME				
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.71	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
MASS				
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
TEMPERATURE (exact)				
°C	Celcius temperature	1.8C + 32	Fahrenheit temperature	°F
ILLUMINATION				
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
FORCE and PRESSURE or STRESS				
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

* SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.

(Revised September 1993)

Data Model and Relational Database Design for Highway Runoff Water-Quality Metadata

By Gregory E. Granato and Steven Tessler

Abstract

A National highway and urban runoff water-quality metadatabase was developed by the U.S. Geological Survey in cooperation with the Federal Highway Administration as part of the National Highway Runoff Water-Quality Data and Methodology Synthesis (NDAMS). The database was designed to catalog available literature and to document results of the synthesis in a format that would facilitate current and future research on highway and urban runoff. This report documents the design and implementation of the NDAMS relational database, which was designed to provide a catalog of available information and the results of an assessment of the available data.

All the citations and the metadata collected during the review process are presented in a stratified metadatabase that contains citations for relevant publications, abstracts (or previa), and report-review metadata for a sample of selected reports that document results of runoff quality investigations. The database is referred to as a metadatabase because it contains information about available data sets rather than a record of the original data. The database contains the metadata needed to evaluate and characterize how valid, current, complete, comparable, and technically defensible published and available information may be when evaluated for application to the different data-quality objectives as defined by decision makers. This database is a relational database, in that all information is ultimately linked to a given citation in the catalog of available reports. The main database file contains 86 tables consisting of 29 data tables, 11 association tables, and 46 domain tables. The data tables all link to a particular citation, and each data table is focused on one aspect of the information collected in the literature search and the evaluation of available information.

This database is implemented in the Microsoft (MS) Access database software because it is widely used within and outside of government and is familiar to many existing and potential customers. The stratified metadatabase design for the NDAMS program is presented in the MS Access file DBDESIGN.mdb and documented with a data dictionary in the NDAMS_DD.mdb file recorded on the CD-ROM. The data dictionary file includes complete documentation of the table names, table descriptions, and information about each of the 419 fields in the database.

INTRODUCTION

Nationally, literature searches to obtain existing information and data relevant to studies of highway water-quality use a substantial portion of research funds at the State and Federal level (Transportation Research Board, 1997). Knowledge of existing information and expertise may be of great value to researchers and decision-makers. Having this information may facilitate enhancement of existing knowledge rather than repeating efforts when evaluating the characteristics of highway-runoff water quality and the potential effects, and mitigation of highway-runoff constituents on water quality and ecosystems in receiving waters. Knowledge of the existing literature also may provide information necessary to address regulatory issues such as for Non-Point-source Discharge Elimination System (NPDES) permits (Swietlik and others, 1995) or for assessments of total maximum daily loads (TMDLs) in receiving waters potentially affected by highway runoff discharges (Rossman, 1991). Therefore, a national database of information relevant to the study of highway-runoff quality is necessary to facilitate and optimize future research (Transportation Research Board, 1997). The Federal Highway Administration (FHWA) and the U.S. Geological Survey (USGS) cooperated on the National Highway Runoff Water-Quality Data and Methodology Synthesis (NDAMS) to catalog available information, to assess a sample of this information, and to document results of the synthesis in a format that would facilitate current and future highway and urban runoff research. The existing literature was examined to determine if it is valid (useful for intended purposes), current, and technically defensible.

The NDAMS database is a relational-bibliographic database in which all information is linked to a citation in the catalog of available reports. Entries in all data tables are linked to a particular citation, and each data table focuses on one aspect of the information collected in the literature search and evaluation of available information. The targeted users of NDAMS metadatabase are Federal and state highway engineers and decision makers, regulators, USGS personnel, and others interested in the quantity and quality of highway and urban runoff. This database is implemented in the Microsoft (MS) Access database software because MS Access is widely used within and outside of government and is familiar to a variety of existing and potential customers. In addition, it can be used to export custom data sets in a variety of formats, and it provides a direct migration path to larger database management systems [for example, MS Structured Query Language (SQL) Server or Oracle] and Internet data access (for example, Reece and others, 1999). The Access database environment provides a strong relational database management system (using standardized SQL constructs, data types, and built-in integrity tools) and flexible access to the data. Desired linkages to other databases (geographic information systems (GIS), water-quality databases, and bibliographic information sources) can be accommodated with extensions or minor modifications to the NDAMS model.

This report provides basic information that will help decision makers understand the design, functionality, and potential uses of the NDAMS stratified metadatabase. The report describes the NDAMS logical data model and its physical implementation as an MS Access database for knowledgeable users who may need to do one or more of the following:

- obtain an introduction to the NDAMS data model;
- evaluate the NDAMS storage structure for a specific use;
- customize a copy of the NDAMS database;
- link this database to other types of databases (such as GIS);
- build software applications servicing the NDAMS database for data entry, data exploration, or reporting.

The intended audience for this document will have some background in the design or use of relational databases. Information about data models and relational database design concepts are available in many books (for example, Fleming and von Halle, 1989; Hernandez, 1997; Roman, 1999), and in the Federal data modeling standard document FIPS 184 (National Institute of Standards and Technology, 1993) for potential users who may need to become acquainted with these topics.

DATABASE DESIGN AND IMPLEMENTATION

The database is designed to store large amounts of information in a structure that allows a practitioner to explore, synthesize, and review related metadata efficiently. The NDAMS database is designed to be rigid with regard to protecting the integrity of the data once entered, yet it is flexible enough to store variable data, fully open to extension and customization, and able to provide linkages to other databases. A normalized relational design underlies the physical database, and built-in data integrity and quality control features common to modern database management applications, such as MS Access, are used to fully guarantee the characterization and reliability of each data element.

All the citations and the metadata collected during the review process are presented in a stratified metadatabase (fig. 1). The database is stratified because

- it is populated with as many relevant bibliographic references as obtainable,
- it contains an abstract for reports sponsored by government organizations,
- it contains an abstract (or previa) and review information for the references that were reviewed, and
- it contains the result of metadata reviews (Dionne and others, 1999) for a sample of selected reports that document the results of highway or urban runoff investigations.

The database is referred to as a metadatabase because it contains information about available data sets rather than a record of the original data. The database contains the metadata needed to evaluate and characterize how valid, current, complete, comparable, and technically defensible the published and available information may be when evaluated for application to different data-quality objectives as defined by decision makers concerned with the quality of and potential ecological effects of highway runoff. The potential utility of available data is based on the current understanding of the technical concerns and documentation required to substantiate verifiable data (Granato and others, 1998).

The purpose and scope of the project, the design of the metadata review process, and potential perceptions of customers mandated a relational database structure. The purpose and scope of the project was to catalog the existence of as many published and available reports as possible that are relevant to the study of stormwater-runoff. Only abstracts that do not have copyright restrictions (or previas for reviewed journal articles)—about 50 percent of the reports cataloged—were to be included with the database. The metadata review process was intended to evaluate a sample of the available population—about 10 percent of cataloged reports. In each metadata review, only the information pertinent to the subject report would be included in the review (Dionne and others, 1999). For example, if the report being reviewed did not contain stormwater-flow data, it would not be appropriate to include information on flow-monitoring methods for the subject report. Because these factors would produce a stratified metadatabase with many records of general information at the top (the citation level) and relatively few records, but very detailed information (many fields) at the bottom, it would be very inefficient to take a one-table spreadsheet approach. Using one (or a few) general tables would obscure the organization of the metadata and would make comparisons difficult because the information from the metadata reviews would be diluted among a sea of null entries for particular citations whose specific metadata details were not available. In the relational database design, however, a number of topic-specific data tables containing only records for citations in which specific metadata was documented indicate more clearly the information that is (or is not) properly documented in the relevant literature.

Model Specifications

Data and user requirements are a necessary prerequisite for any database development effort. These requirements are used to clarify the purpose and scope of the database, to guide the development of the database structure, and to serve as final checkpoints for the

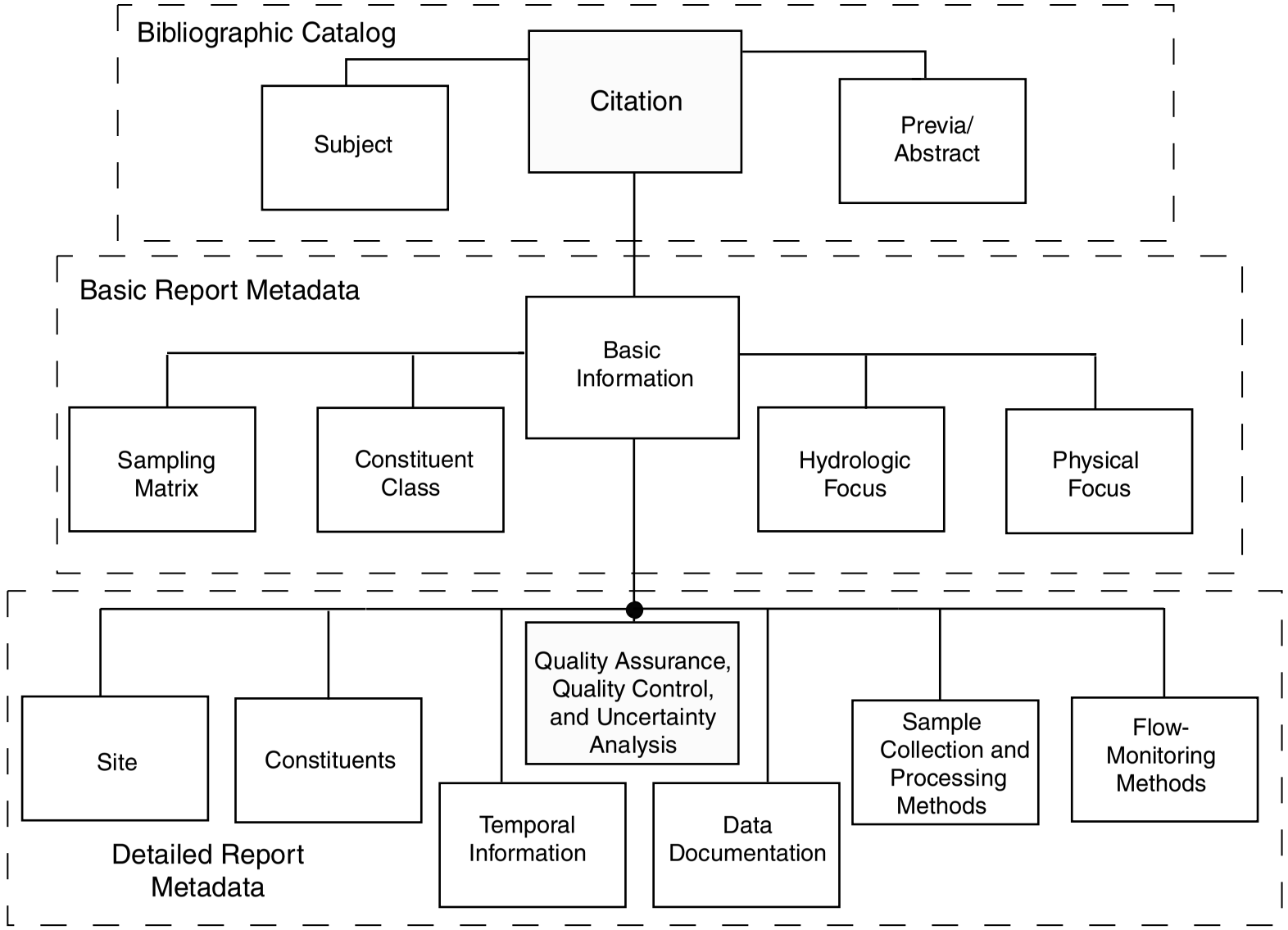


Figure 1. Schematic diagram of the organization of the stratified metadatabase.

end product. The NDAMS database is designed to meet the need for improved information exchange by providing a database structure that can be used to

- catalog available publications relevant to the study of highway runoff quality;
- provide the metadata necessary to evaluate if the existing information is valid, current, and technically defensible for a given problem with known data-quality objectives (Granato and others, 1998);
- provide a stratified metadatabase that will facilitate current and future information needs by providing bibliographies, abstracts, and (for select reports) metadata; and
- provide a logical database design that will facilitate future extension when suitable data sets are incorporated into a National Highway Runoff Water-Quality Database.

Software Specifications

Many issues were considered in choosing database software, including import and export capabilities in different formats, reasonable purchase price, prospects for continued availability, software capabilities, ease of use, and vendor support. A primary consideration is to have software that is user-friendly with an interface that would be familiar or easy to use for many potential customers. Import and export capabilities, including the ability to read and write text and data in space-delimited or tab-delimited ASCII formats, are important to facilitate information transfer between text files, word processors, spreadsheets, and other database applications. The interface must support input, output, and queries through standard forms for use by report reviewers and database users. It is also important to be able to exchange data and information with GIS software (such as Arcview) to facilitate geographic data analysis. Additionally, the database software was chosen to allow data to be made available through the internet/world-wide web without a large investment in additional equipment, software, or a computer specialist's time. Logistically, it is also important to select database software that

- can be published on CD-ROM with a low/no royalty run-time version for basic use by people who do not own a particular database-software package;
- has a reasonable purchase price for advanced users who may not own the software;
- is available for IBM-type personal computers, which are predominant among Federal and State Department of Transportation (DOT) offices; and
- has vendor support from a company that has a reasonable chance for continued viability over the next 3 to 10 years.

Additionally, the ability to export text files that contain bibliographies in a Government Printing Office style (including appropriate punctuation) is important for this bibliographic package. MS Access was chosen from the several possible database systems because it best fit these design criteria and because it is the same software that is being used to support the FHWA ultra-urban best-management practices database (Cazenas, 1998; Shoemaker and others, 2000).

The Modeling Process

Creating a model that accurately describes the structure of a set of data is a precursor to any correctly designed database (Fleming and von Halle 1989). Logical modeling is the process of analyzing and reducing a set of data to its separate components, establishing the nature and direction of relationships[1] between those components, and thereby building a structure for the data that automatically enforces the business rules that are needed to maintain data integrity and provide easy access to all of the information stored in the database. The NDAMS database began as the sum of the metadata requirements defined by subject matter experts in the USGS, the FHWA, and various state DOTs. These requirements were codified into a set of report review protocols documented by Dionne and others (1999).

A properly designed logical-data model provides for change in the data management process. A logical-data model should therefore represent a straightforward generic-structure that easily allows for later corrections and extensions without affecting the integrity of data

[1]In this report, the term relationship refers to an association between two entities or between instances of the same entity (National Institute of Standard and Technology, 1993).

already residing in the database. Furthermore, a logical data model is used to create and maintain a data dictionary of all the elements in the database, and thus provides a common language and reference for all users.

In general, a logical data model can be discussed in grammatical terms. Entities (tables) in a data model can be thought of as nouns, their descriptive attributes (fields) can be thought of as adjectives that describe each instance of a noun, and the relationships between entities can be thought of as verbs that clearly and concisely define the nature and restrictions of the associations between nouns. An example from the NDAMS database might be that "a Citation may have more than one Study Site" where the nouns identify two tables (Citation and Site), and the verb describes a one-to-many relationship between them (each Citation has associated data for one or more Study Sites).

All the information needed to provide answers to questions about available publications and the quality of the existing highway-runoff data set as defined by Granato and others (1998) and Dionne and others (1999) should be present in the database. These items should be present in the model in a form that is familiar to the user, and wherever possible, in terms that are well-defined and in common usage among those familiar with highway and urban runoff-quality issues. A correctly designed database, based on a robust logical model, includes methods to ensure that the data in the database meet a predetermined level of detail and accuracy. A robust database design also is necessary to provide for expansion as the need for data and information changes with time.

The logical model and the database design also should meet programmatic needs. The NDAMS database was designed to facilitate the project quality assurance and quality control (QA/QC) program and to provide the most comparable metadata from which to assess the existing literature on highway-runoff. To work from a position of experience and knowledge, the design of the NDAMS database was completed when the literature review process was about 75 percent complete. The experience and knowledge of the reviewers was used to translate the results of the review process using the NDAMS review sheets into standard responses that would lead to consistent and objective interpretations of available data in published reports. Whenever possible, narrative descriptions in the review sheets were condensed into yes/no questions (or yes/no variants that would include responses for not applicable or unknown where appropriate), standard multiple-choice questions, or extendable lists of appropriate responses. Information in the review sheets was converted to standard inputs in the database by using standard lists in domain tables to ensure that the data could be reliably and repeatably entered, and that information recorded in the database could be grouped and (or) classified for interpretation. For example, the reviews include information about sampling materials (such as equipment, bottles, and preservatives), but examination of the reviews indicates that this information was not typically available, or was not described in a consistent manner among the reports reviewed. Therefore, the question regarding sampling materials was simplified to a yes or no response. Also, some information from the review forms was omitted when it was deemed that inclusion in the database would not add meaningful information. For example, although it is recognized that trained and professional sampling teams are necessary to collect reliable stormwater-quality data in a consistent manner, it proved impossible to quantify this measure of data quality in an objective review of a published report, and therefore, this information was omitted from the database.

Once designed, the logical data model was used to create the physical implementation of the database by providing table and field names, their discrete attributes (for example, field type, size, and definition), their relationships, and the complete key structure needed to ensure data integrity. The NDAMS database was designed using the data-modeling tool ERwin version 3.5 (Platinum Technologies). Database modeling tools facilitate rapid design and modification and invisibly handle many important details that could affect the quality of the data structures and individual elements. Many of these tools also allow a single data model to be used to generate a physical database in several different database-management systems (MS Access, Oracle, DB2, Ingress, and others). For a given target database, ERwin provides full access to application-specific data element properties and control over the creation of rules and constraints that guarantee the integrity of the relational design and the data that it is intended to contain.

Design Conventions

Several design conventions were adopted for use in the NDAMS database. These conventions provide standardized rules for creating and assembling the various parts and frame communication about how the different elements of the database function together.

Normalization

The NDAMS model is structured to approximately "third normal form." The result of this is that each table stores information dependent on a single key, each datum or descriptive element is only recorded in one location, keys to those data are dispersed among other tables as a surrogate and entry point to the detail data, and all information in the database is accessible from anywhere else through the use of the key relational structure. Custom queries can create the illusion of a flat file (or spreadsheet) view that combines many separate related tables at once for data entry, review, or reporting, while the correct data structure is hidden from the user. MS Access provides direct protection of the keys throughout the tables in the database, and does not allow the key structure or defined relationships to be corrupted or compromised; this is one part of maintaining referential integrity of the data.

Keys and Relationships

Each table in the database is designed to hold a specific, defined, and delimited set of data, where each record in the table (a row, or the particular collection of data it represents) is identified by a unique primary key (PK) value, a number. The primary key and its value in one table can be used to identify the records in another table that has related information. For example, a unique citation identification number (Citation_ID) identifies each report in the database; this key is used to identify all information in the database that is associated with any given report. In the relationship, the PK from the parent table is called a foreign key (FK) in the child, or receiving table. The relationship established between tables through the shared PK/FK fields provides a path for the data to be easily combined into a single view of the records in the two or more source tables. For example, the Citation_ID is the PK in the citation table and the FK in the abstract table, and this relationship identifies each abstract with the appropriate citation. With many tables related to one another in various combinations, the relational model provides a way to extract data from any two or more tables in any combination desired, if an unambiguous pathway of relationships exists between them. The accuracy of data replicated through the key when it is viewed or combined with data from other parts of the database is controlled and ensured by limiting the field complexity of the data stored in individual tables. Thus information-rich keys (fields that have apparently unique values, such as a code name, and could serve to identify data rows) are not allowed to be key fields, but are used as attributes identified by an information-neutral numeric key. The possibility that information-rich key fields could be reassigned in practice and thus corrupt associated data in other tables was reason enough to prohibit their use as a PK in the NDAMS model. In some tables a complex key consisting of two or more fields together could serve as the PK (for example, the combination of citation identification and a site number in the location tables), but the convention used in NDAMS database, where possible, was to relate tables by providing a single field to serve as the surrogate for any such complex PK (association tables excepted).

The MS Access database software protects PK fields in two ways: (1) it requires that the PK field always be filled whenever new rows are added, and (2) it prevents the PK field value from being duplicated in rows of the table (each PK value within a table is unique within that table). This is another part of maintaining referential integrity of the data. Most data tables in the NDAMS database have a single field that serves as the PK, and because this is a bibliographic database and all information is ultimately linked to a particular published report, the key is often the citation number. In tables that supply lists of choices, the key itself is often simply an incremental integer to represent each row of data in the table. Exceptions are the association tables described below, where a single key field is not practical or desirable.

Relationships between tables using keys are modeled and built into the database. The MS Access database software enforces the defined-relational properties. To protect the integrity of the keys across tables, referential integrity constraints are placed on the keys through the relationship. All relationships in the

NDAMS database have common settings to protect the keys. No keys are optional; each record in each table must be identified by a key value. A cascade update applies when a PK value is changed; all occurrences of that value in the related FK in other tables have their values automatically updated. A restrict delete applies when a record is being deleted from a table and prompts a warning when children of the record (holding that FK value) are present in other tables, thus preventing related records from becoming orphaned. MS Access does this warning automatically once the rule is established. Relational rules can be adjusted as needed after experience with the NDAMS database, but these rules, along with enforcement of the uniqueness of PKs, provide the basic protection needed to prevent key corruption and maintain referential integrity throughout the database.

Definition of the NDAMS Core Logical Entities

All the citations and the metadata collected during the review process are presented in a stratified metadatabase (fig. 1). The top level is a bibliographic-catalog database that provides information about relevant, available, and published literature. The second level catalogs general metadata for all reports reviewed. The third level catalogs detailed metadata for reports that document results of water-quality investigations. Report reviews are divided between two categories—review/summary reports that provide a limited amount of metadata, and data/interpretive reports that provide detailed metadata (Dionne and others, 1999). The core logical entities are (1) the bibliographic (citation) database, (2) the basic report metadata, and (3) the detailed report metadata. In the bibliographic catalog database are the citations, the subject, and when applicable, an abstract. In the next level of the stratified metadatabase are the basic report information, sampling matrix, constituent class, hydrologic focus, and physical focus, each representing a core logical entity. The core logical entities for site information, chemical constituents, temporal information, quality assurance, quality control and uncertainty analysis, data documentation, sample collection and processing methods, and flow-monitoring methods are in the detailed report metadata. One or more tables in the physical data model represent each of these core logical entities.

The stratified metadatabase designed for the NDAMS program is presented and documented using two MS Access files recorded on the CD-ROM. The file DBDESIGN.mdb is the published version of the stratified metadatabase, and the file NDAMS_DD.mdb documents this design. The database is a relational database in which all information is ultimately linked to a given citation in the catalog of available reports. The main database file contains 86 tables including 29 data tables, 11 association tables, and 46 domain tables. The data tables all link to a particular citation, and each data table is focused on one aspect of the information collected in the literature search and evaluation of available information. The association tables resolve potential many-to-many relationships by pairing records from two or more tables to form valid cominations. For example, an association table is used to pair citations with the months in which sampling efforts were conducted, thus allowing each citation to be associated with more than one month, and each month to be associated with more than one citation. The domain tables provide lists of acceptable standard choices for data entry, organization, and interpretation.

The NDAMS data model makes extensive use of lists for classification or description of items in the database. These are called domain tables because each provides specific information about a particular subject domain. About 53 percent of the tables in the database are domain tables that serve the main data tables. An example of a domain table is tdxState, where each record is for an individual State or province (for foreign countries), and the fields provide kinds of information about each State (for example, name, postal abbreviation, code). A relationship with a domain table can therefore provide many different pieces of information through the single value of the inherited key field, and also significantly enhances sorting and grouping options for the related data by providing discrete tables where such actions can be performed.

Domain tables provide additional power to NDAMS database because they can be extended at will, both structurally (by adding new descriptor fields) and list-wise (by adding to the list), without affecting the rest of the database design. The value of the many separate domain tables needs to be emphasized because

it is also possible to include some domain table fields in the data table that uses that domain. Normalization rules, however, dictate that data be compartmentalized as much as possible to avoid and prevent redundancy in the data. Use of domains also reduces the chance for non-unique data input. For example, if a state is entered directly in a data table as a text field, it may be entered in many ways (for example, Commonwealth of Massachusetts, Massachusetts, Mass., or MA) and may be misspelled in any of a number of ways. Therefore, standard choices provided by the domain tables ensure proper spelling and prevent misinterpretation caused by inappropriate variations in input choices.

Naming Standards

Naming standards are an important database design tool (Fleming and others, 1989; Hernandez, 1997). They can convey important information about an item's identity or contents—or both—with little more than an understanding of the convention used. In the NDAMS database, naming standards apply to tables and fields.

Name Construction

The three general name construction standards are listed here with explanations.

1. Names are constructed using whole words (with a few exceptions) to provide as much intuitive meaning as possible.
2. Names that are composed of word phrases have the first letter of each word in the phrase capitalized to assist in reading the name.
3. Names are constructed of only alphabetic characters, with no numerals, punctuation or other special characters (such as spaces or dashes) allowed, except for primary keys which consist of the table name with an _ID suffix, and association tables which have names assembled from the parent tables separated by underscores.

Examples of phrased field names include Authors, YearOfPublication, AuthorAffiliation, and PurposeDescription. Names that do not use whole words in their composition are constructed using an abbreviation [such as YNNA (for the choices Yes, No, or Not Applicable), YNU (for the choices Yes, No, or Unknown), NCME (for the choices No, Calculated, Measured, or Estimated)], or a commonly accepted acronym or abbreviation, specifically BMPUsed (for best management practice used).

The exception to the no-punctuation rule applies only to some PK fields and association tables. The PK of many data and domain tables is the table root name with a suffix of _ID, such as State_ID as the PK for the table tdxState. This standard of using an _ID suffix for key fields allows visual distinction between key and non-key (data) fields. The visible presence of _ID keys as non-PK fields in a table also clearly indicates that one or more data fields are available in a related table of the same name.

Table Functional Prefixes

In the NDAMS database the naming standards for tables are a modification and extension of those in common use by Access developers and MS Visual Basic programmers. To facilitate rapid identification and communication of the function of different tables in the database, seven functionally distinct types of tables are defined by a three-letter prefix applied to the table name. Therefore, table names also are distinguishable from field names because tables have lower case prefixes and fields do not. The seven (three-letter) table-name prefixes used in NDAMS and their definitions are as follows:

tbl–simple data table, for example tblCitation. These tables hold primary data entered into the database. The data table tblCitation has a single PK field, which is an integer automatically incremented for each new record (an AutoNumber field in MS Access) and is guaranteed to be unique through its creation by Access itself. Other tables typically contain one or more fields that are FKs from domain lists, and thus inherit data elements from other tables as their own extended attributes.

tas–simple association table, for example tas_tblTemporalInformation_tdsMonth. These association tables are used to resolve many-to-many relationships between other tables and are composed of only key fields. In the example of tas_tblTemporalInformation_tdsMonth, constructing an association table allows a citation to have more than 1 month sampled, or a single month sampled to be identified for a number of different citations. The association table name is a phrased combination of the two

(or more) parent table names sharing in the association (here, tables tblTemporalInformation and tdsMonth). They have a complex PK composed solely of the FKs of each table involved in the association (here, Citation_ID and Month_ID); no other fields are present in tas tables. The PKs in association tables are assembled by deliberately pairing records from the parent tables in the desired combinations. MS Access maintains the integrity of the relationship by allowing each combination only once, thus ensuring the primacy of the key.

tad–association table with data, for example, tad_tblTemporalInformation_tdsStormType. This type of table has the same properties and functionality as tas association tables, but also stores data, which describes some property unique to the association itself (that is, the combination of key values forming the PK). For example, tad_tblTemporalInformation_tdsStormType combines a Citation_ID with a particular StormType_ID to form its PK. This allows individual citations to be associated with more than one storm type, and each storm type to be associated with more than one citation. This table also adds a non-key field, NumStorms, which defines how many of each of the storm types are represented by the pairing of an individual citation with a specific storm type. Thus, tad tables resolve many-to-many relationships and also store data unique to a specific association of items from different tables that cannot be realistically stored anywhere else.

tds–static domain table, for example, tdsSubject. These tables typically hold a list of classification or descriptive items that are used by other tables; the list of choices is considered static in that its contents are believed to be complete (for example, the tdsSubject table is fully populated with records describing all the keywords that were selected for classifying reports in the database). These are also sometimes known as lookup or reference tables. They have a single PK field that is a simple integer and the key must be manually entered whenever new entries to the list are made, thus assuring that additions to the static list are not done inadvertently. Manual key entry also allows creation of custom unique key values for use with the choice list, rather than being limited to automatically generated numbers (for example, adding a 0 key value to indicate not applicable). The tds domain tables also include two flat files denoted by an FF suffix, in which an organizational hierarchy in a related recursive domain is expanded so the user can better observe the nesting of elements in the recursive tables.

tdx–user-extendable domain table, for example tdxCountry. These are domain tables that provide a list that the user will add to as needed, and thus differ in an important way from tds tables. The tdx tables are considered incomplete at the outset and cannot be populated fully in advance of use of the database. For example, the tdxCountry table will have new entries whenever new countries are identified in reports that are reviewed. The tdx tables have a single PK field that is an AutoNumber field (automatically incremented integer) to allow some automation of additional entries without regard to key value assignment. Although similar in construction to tbl tables, tdx tables function to provide an extendable domain list for characterizing records in data tables.

tdm–multiuse domain table, for example tdmYNU. These tables typically hold a list of choices that can be used as the selection for a number of different fields. For example the domain table tdmYNU provides the choices Y (yes), N (No), and U (unknown) in 14 fields distributed among 6 different tables. In each case where tdm domains are used, the receiving field has a different name (it is role-named; for example, tdmYNU is used by the fields PeerReview, DataAvailable, and others in table tblBasicReportInformation). This is one of the characteristics that make the tdm table different from the other domain tables. Multiuse domain tables also have no numeric key and are strictly used to represent the list of valid choices for two or more fields in the database. The list of choices is considered static in that its contents are fixed because more than one field in the database uses the options included. If there is a need to modify the available choices in one field, a new tdm table should be established to provide the necessary choices, unless the addition is considered valid for all uses of the tdm table.

tdr–recursive domain table, for example tdrConstituent. These tables also are domain tables in that they hold a list of classification or descriptive items that are used by other tables to provide a list of choices. These tables, however, provide the given choices within an organizational hierarchy by recursive association. For example, tdrConstituent includes records for

both constituent identity and the classification terms used to group constituents. Each record can point to a parent record in the same table to establish the hierarchy of terms, and thus recursively defines each constituent within the hierarchy of the water-quality constituent classification system described by Dionne and others (1999). The relationships between constituents and constituent classes are defined in the table in a manner that facilitates data manipulation in the MS Access environment as denoted by the self-join relationship on the recursive tables in the relational database design diagram. To assist the user with understanding the hierarchy of terms, the two tdr tables (tdrConstituent and trdMatrix) also are represented as tds flat files with a FF suffix (tdsConstituentFF and tdsMatrixFF).

With practice, the naming conventions described above can become a useful feature in exploring the NDAMS database. Creation of new fields or accessory tables to extend the functionality of the database should use these naming conventions to ensure uniform communication about the functional properties of the new items.

Entity/Relationship Diagramming Conventions

Entity/Relationship (E/R) diagrams are used to visualize database designs. Several different display and notation methods are in common use for E/R diagrams, but all share similar characteristics. The Information Engineering (IE) relationship notation and style provided by ERwin (and enhanced by NDAMS naming standards) are used here. An E/R diagram that illustrates several diagramming conventions is shown in figure 2. In this E/R diagram boxes are used to denote an entity, which equates to a single table in the physical database. Each entity box has its name at the top. Within the box are one or more entity attributes, which equate to fields in the physical database. Lines connecting entities represent their defined "relationship" and are often called joins in the physical database.

In the E/R diagram, the PK for each entity (composed of one or more attributes) is listed at the top of the attribute list within the entity box and is separated from the other attributes by a horizontal line. When a PK from one table (the parent) is passed or migrated to another table (the child) through a relationship, this FK in the child table has the FK designation in the diagrams. If the FK is part of the primary key in the child table, the relationship is said to be strong and the relationship line is solid in the diagram (this will always be true for association tables). If the FK is a non-PK attribute of a child table, the relationship is said to be weak and the line is dashed. To further help visualize table dependencies, tables in a strong relationship (FK is part of the PK) are shown with rounded corners, whereas tables that do not have FK dependencies in their PK are shown with squared corners.

Each relational line has a direction and cardinality. The direction is recognized by the origin end (parent entity in the relationship) usually having a single line perpendicular to the relationship line, whereas the target end (child entity in the relationship) shows a more complex cardinality symbol—usually a branch. Relationship cardinality symbols are the shorthand for whether the relationship is one-to-one (1:1, each entity instance in the parent table has exactly one match in the child table), one-to-many (1:n, each entity instance in the parent entity has one or more matching instances in the child entity), or some other variant. The most common relationship in the NDAMS model is a one-to-zero-one-or-many (1:0,n) which is recognized as a child-end symbol consisting of an empty circle with three-branches (a crows' foot). This type of relationship, illustrated in figure 2, is used when each parent can have none, one, or more children, and each child must have a parent (the FK cannot be null). Thus, each CitationType record can serve 0, 1, or more Citations, and each Citation must have a CitationType.
The zero part of the relational cardinality allows a CitationType record to exist in the domain without actually being used until a Citation needs it. This is very useful for filling a domain table with a list of all permissible values before other data are entered into a database. The only other cardinality type in NDAMS is one-to-zero-or-one (1:0,1) where the parent may have zero or only one record in the child table. This is represented by the relationship between tblCitation and tbl Abstract in figure 2, where each citation may have a single abstract available in the tblAbstract table, but not all have abstracts.

Figure 2 presents a detailed view of the bibliographic portion of the database and demonstrates many of the diagramming conventions mentioned. For example, the table tblAbstract is a child (and therefore a dependent table) of tblCitation. There is a one-to-zero-or-one relationship (each citation may have only one abstract), and the Citation_ID from

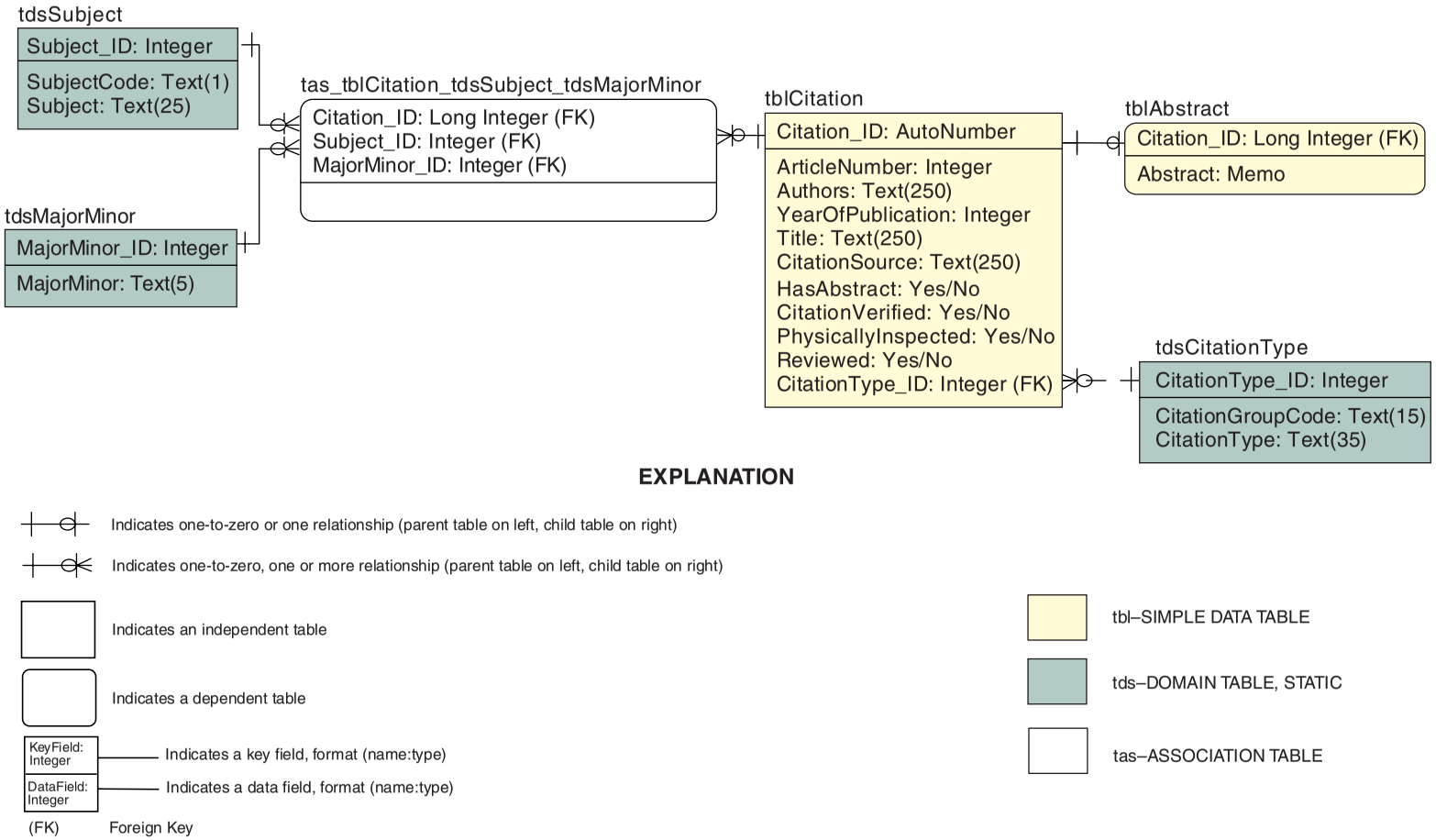


Figure 2. Example of an entity/relationship diagram of the bibliographic portion of the database.

tblCitation is the FK that identifies each abstract as being related to a particular citation. The table tdsCitationType, however, is an independent table supplying choices to tblCitation. This is a one-to-zero-one-or-more relationship (each report type in tdsCitationType may be associated with a number of reports in tblCitation) with tblCitation being a child (but nondependent) and receiving the FK CitationType_ID from tdsCitationType. The association table tas_tblCitation_tdsSubject_tdsMajorMinor is used to group each citation with one-of-many subjects as a minor or major keyword code. This table is therefore a dependent child and is the subject of one-to-many relationships with the citation table and the relevant domain tables. This association table receives a FK from each of the parent tables as indicated in figure 2.

DESIGN DOCUMENTATION

Design documentation facilitates the current use and potential modification for future use of the database. The NDAMS database contains 86 tables, consisting of 29 data tables, 11 association tables, and 46 domain tables. The design of the database is fully documented in four ways, including

- documentation in the structure of the NDAMS database file,
- a data dictionary file,
- a detailed database-design diagram, and
- this database-design report.

Each of these products is included in digital format on the enclosed CD-ROM to facilitate use of the database and the underlying design.

The NDAMS database contains design documentation in three formats in the MS Access database. These formats include a relational view of the database tables, table definitions in the properties view, and field details in the table design view. The relationships between database tables may be viewed using the tools menu and selecting the relationships option. The NDAMS database, however, is complex enough to limit the clarity of information available in this view. Each table in the database is defined in the table description window of the properties feature of each table (viewed by right-clicking on the table and selecting properties from the ensuing menu pop-up window). Each field in all the tables is defined in the table structure. Field definitions can be read in the status box at the lower-left of the MS Access interface screen when a table is open in datasheet view. Field details are available in table design view, including

- the primary key field(s) (as denoted by a key symbol),
- field names,
- data types,
- description (field definition), and
- specific field properties.

The table-design view allows the reader to assess and manipulate these settings. The documentation provided for objects in MS Access is useful but does not provide an overview or illustration of the overall design of the database.

A computerized data dictionary file is provided to facilitate examination of the database design and for documentation of the completed database. Complete documentation of the table names, table descriptions, and information about each of the 419 fields in the database is provided in the data dictionary file**,** which is also an executable MS Access file. The data dictionary file is named NDAMS_DD.mdb, and this file also is available on the CD-ROM under database documentation information. The source code for the data dictionary file was developed and distributed as freeware outside the purview of the USGS and is available on the internet (Tessler, 2000). Use of the data dictionary file is described in appendix 1. This file provides summary information about the design and implementation of each table and is very useful for browsing the design of the database. The data dictionary file, however, does not provide the overview needed to convey the overall design of the database.

A database-design diagram (plate 1; on CD in back pocket) is provided to facilitate examination of the database design and for documentation of the completed database from the perspective of the overall relational structure. Details of the data model are documented on this enclosed database-design diagram. The diagram shows the stratification of the database into the bibliographic, basic report information, and detailed report information. Each table is represented by a rectangle, and individual field information is provided within these units. The primary key fields in the database are in the top division of each table. Relationships between tables are indicated by the con-

necting lines and symbols, which follow the entity-relationship diagramming conventions. Functional table types in the database are indicated by table color, as specified on the plate. The one-field multiuse domain tables are not shown on the diagram because these tables (such as the tdm YNU table) are not important for understanding the overall structure of the database. Information about these domain tables is included in appendix 2 so that the use of these tables will not be misunderstood. Also, the one-field multiuse domain tables are not shown on the diagram because they are used frequently, and including them might impair understanding of the overall structure because of the congestion created by the multiple links. The use of naming standards and functional prefixes becomes apparent as the relationship between tables is illustrated. This database-design diagram is useful for understanding the existing structure and for facilitating future modifications of the database.

SUMMARY

Availability of a database of information on the water-quality of highway and urban runoff may facilitate the enhancement of existing knowledge for evaluating the characteristics of runoff quality, potential ecological effects, and best management practices for mitigating these effects. The FHWA and the USGS cooperated on the National Highway Runoff Water-Quality Data and Methodology Synthesis (NDAMS) to catalog available information, to assess a sample of this information, and to document results of the synthesis in a format that would facilitate current and future research on highway and urban runoff. The existing literature was examined to determine if it is valid (useful for intended purposes), current, complete, and technically defensible. This report documents the design and implementation of the NDAMS relational database, which was designed to provide a catalog of available information and the results of an assessment of the available data.

All citations and metadata collected during the review process are presented in a stratified metadatabase. The database is stratified because it is populated with as many relevant bibliographic references as obtainable; it contains an abstract or a previa for a subset of these reports, and contains metadata for a subsample of reports that document the results of highway- or urban-runoff investigations. The database is referred to as a metadatabase because it contains information about available data sets rather than a record of the original data. The database contains the metadata needed to evaluate and characterize how valid, current, complete, comparable, and technically defensible published and available information may be when evaluated using different data-quality objectives, as defined by decision makers concerned with the quality of and potential ecological effects of highway runoff. The potential utility of available data is based on the current understanding of the technical concerns and documentation required to substantiate verifiable data.

This database is a relational database in that all information is ultimately linked to a given citation in the catalog of available reports. The main database file contains 86 tables consisting of 29 data tables, 11 association tables, and 46 domain tables. The data tables all link to a particular citation, and each data table is focused on one aspect of the information collected in the literature search and evaluation of available information.

This database is implemented in the MS Access database software because it is widely used within and outside of government and is familiar to a number of existing and potential customers. The stratified metada-

tabase designed for the NDAMS program is implemented in the MS Access file DBDESIGN.mdb and documented using the NDAMS_DD.mdb data dictionary file recorded on the CD-ROM report. The data dictionary file includes complete documentation of the table names, table descriptions, and information about each of the 419 fields in the database.

REFERENCES

Cazenas, P.A., 1998, Ultra-urban best management practices: Federal Highway Administration Fact Sheet, 4 p.

Dionne, S.G., Granato, G.E., and Tana, C.K, 1999, Method for examination and documentation of basic information and metadata from published reports relevant to the study of stormwater runoff quality: U.S. Geological Survey Open-File Report 99-254, 156 p.

Fleming, C.C. and von Halle, Barbara, 1989, Handbook of relational database design: Reading, Mass., Addison-Wesley Publishing Company, 605 p.

Granato, G.E., Bank, F.G., and Cazenas, P.A., 1998 Data quality objectives and criteria for basic information, acceptable uncertainty, and quality-assurance and quality-control documentation: U.S. Geological Survey Open-File Report 98-394, 17 p.

Hernandez, M.J., 1997, Database design for mere mortals—A hands-on guide to relational database design: Reading, Mass., Addison-Wesley Publishing Company, 480 p.

National Institute of Standards and Technology, 1993, Standard for integration definition for information modeling (IDEF1X): Federal Information Processing Standards Publication 184, 155 p.

Reece, B.D., Sechen, G.M., Jr., and Granato, G.E., 1999, Search the Federal Highway Administration National Data and Methodology Synthesis Bibliography: accessed on October, 16, 2000, at URL http://ma.water.usgs.gov/FHWA/biblio/.

Roman, Steven, 1999, Access database design and programming (2d ed.): Sebastopol, Calif., O'Reilly and Associates, 409 p.

Rossman, L.A., 1991, Computing TMDLs for urban runoff and other pollutant sources: U.S. Environmental Protection Agency Final Report EPA 600/A-94/236, 17 p.

Shoemaker, L., Lahlou, M., Doll, A., and Cazenas, P., 2000, Stormwater best management practices in an ultra-urban setting: selection and monitoring: Federal Highway Administration Report FHWA-EP-00-002, 287 p.

Swietlik, W.F., Tate, W.D., Goo, R., and Burneson, E., 1995, Strategies for using NPDES storm water data, *in* Torno, H.C., ed., Stormwater NPDES Related Monitoring Needs: Crested Butte, Colo., American Society of Civil Engineers, August 7–12, 1994, p. 244–276.

Tessler, Steven, 2000, DataDict—Two minutes from your Access 97 database to a printable data dictionary: accessed on October, 31, 2000, at URL http://www.CleanDataSystems.com/.

Transportation Research Board, 1997, Environmental research needs in transportation: Transportation Research Board, National Research Council, Washington, D.C., Circular, no. 469, 98 p.

Appendix 1: Use of the Data Dictionary

Relevant information for documenting the database was compiled into an electronic data dictionary in Microsoft Access, NDAMS_DD.mdb.

1. Download the file from the CD-ROM,
2. Reset the file properties so that the file is not read-only (right click on the file icon on your drive, choose Properties and ensure that the read-only box is not checked, click ok), and
3. Double click on the interface to activate the MS Access window.

Opening the file displays the data dictionary main menu.

This menu provides an interface to

- preview or print a report,
- browse table and field information,
- examine table information, and
- examine field information.

The "preview or print a report" button on the main menu activates a formatted MS Access report that details table and field information from the database. This report is comprehensive with a summary page and a number of subsequent pages, each of which represents the design information for an individual table in the NDAMS metadatabase. This report may be explored in this preview interface and printed.

The "browse table and field information" button on the main menu activates a browse interface that provides summary information about the total number of tables, fields, and records in the database. This form provides a browse window from which any table in the database can be selected for exploration of detailed design information. This interface also contains a data window in which the details for the selected table are presented.

The "examine table information" button activates a menu that allows the user to view file information by type of table. This menu includes data tables, domain tables, and association tables. Each of these choices provide the table names, the number of fields in each table, and the table description for each table in each respective group of tables.

The "examine field information" button activates a menu that allows the user to view information for the different fields. This menu allows the user to examine

- field types by table,
- fields used in more than one table,
- all fields listed alphabetically,
- key field properties, and
- non-key field properties.

The "field types by table" button provides a list of all tables, the number of fields in each table, and the type of data (Date/Time, Memo, number (integer), number (long), number (single), text, or Yes/No). The "individual fields used in more than one table" button provides a list of field names and the number of uses (within a table and (or) in different tables) for that field. The "fields listed alphabetically" button provides a list of all fields with the number of tables in which the field is used and a description of the field. The "key field properties" and "non-key field properties" buttons provide a field description for each respective field.

Appendix 2: Multiuse Domain Table Implementation

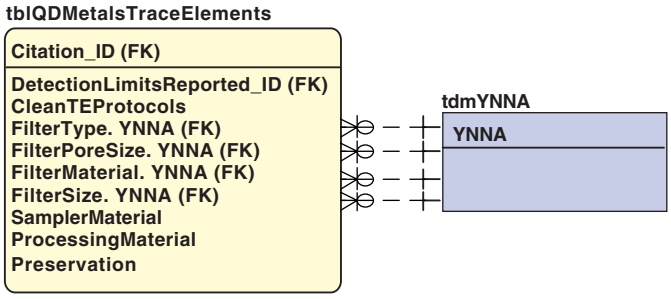
Multiuse domain tables hold a list of simple choices that can be used as the selection for more than one field. Typically the receiving field poses a question that the domain selection answers. For example, the FilterType field poses the question whether filter type information was provided in a report and is answered by a selection from the tcmYNNA domain: Yes, No, or Not Applicable. Fields using these domain tables can be recognized in the E/R diagram by having a FK designation but without a domain table attached. The multiuse domain tables that are not included on the NDAMS data model diagram are detailed here.

tdmYNNA Domain

The tdmYNNA domain is used in only one table tblQDMetalsTraceElements.

tdmYNNA

YNNA
Yes
No
NA



tdmYNU Domain

The tdmYNU domain is used 14 times in 6 different tables (listed only, not illustrated), and provides the choices Y (yes), N (no), and U (unknown).

tdmYNU

YNU
Y
N
U

Uses are lised below (**table**—fields):

- **tblSiteHighway**—Curbing;
- **tblSampleHandling Methods**—LabCertified;
- **tblQDOrganics**—SampleLocationVolatization, SamplerVolatization;
- **tblQAQCField**—QAQCAudit, PrereleaseDataVerification, and TechnicalMethodsReview;
- **tblQWMethod**: ContinuousLevelMonitoring, ContinuousQWMonitoring, and FirstFlushSamples; and
- **tblBasicReportInformation**—PeerReview, QAQCProgram, DataAvailable, ElectronicAvailability.

tdmNCD and tdmNCME Domains

The tdmNCE and tdmNCME domains are used only in the tblBasicReportInformation table. That table also has tdmYNU domain usage.

tdmNCE

NCE
No
Calculated
Estimated

tdmNCME

NCME
No
Calculated
Measured
Estimated