Note for the monthly list of new publications

Open-File Report 94-327. Design of the Distributed Spatial Data Library (DSDL) for the U.S. Geological Survey, by Douglas D. Nebert. 1994. 30p.

The U.S. Geological Survey has undertaken the design and population of a Distributed Spatial Data Library (DSDL) for use in water-resources investigations. The DSDL includes on-line digital spatial data commonly used in hydrologic applications that apply geographic information systems (GIS) techniques to analyze and display data. The DSDL uses a structured library format and wide-area network access mechanism that facilitates data discovery and exchange among GIS users at more than 80 offices of the U.S. Geological Survey. This report includes information on the structure, content, and software written in support of the DSDL project. A 3-1/2-inch DOS-compatible diskette is provided with the report which contains the DSDL software for use within the ARC/INFO GIS environment under the UNIX operating system.

DESIGN OF THE DISTRIBUTED SPATIAL DATA LIBRARY (DSDL) FOR THE WATER RESOURCES DIVISION, U.S. GEOLOGICAL SURVEY

By Douglas D. Nebert

U.S. GEOLOGICAL SURVEY

Open-File Report 94-327

Prepared in cooperation with the
U.S. ENVIRONMENTAL PROTECTION AGENCY

Reston, Virginia
1994

U.S. DEPARTMENT OF THE INTERIOR
BRUCE BABBITT, Secretary
U.S. GEOLOGICAL SURVEY
Gordon P. Eaton, Director

For additional information
write to:

Chief, Spatial Data Support Unit
U.S. Geological Survey
Water Resources Division
445 National Center
Reston, Virginia 22092

Copies of this report can
be purchased from:

U.S. Geological Survey
Earth Science Information Center
Open-File Reports Section
Box 25286, MS 517, Denver Federal Center
Denver, Colorado 80225

CONTENTS

ILLUSTRATIONS

DESIGN OF THE DISTRIBUTED SPATIAL DATA LIBRARY (DSDL) FOR THE WATER RESOURCES DIVISION, U.S. GEOLOGICAL SURVEY

By Douglas D. Nebert

ABSTRACT

The U.S. Geological Survey is undertaking the design and population of a Distributed Spatial Data Library (DSDL) for use in water-resources investigations. The DSDL includes on-line digital spatial data commonly used in hydrologic applications that apply geographic information systems (GIS) techniques to analyze and display data. The DSDL uses a structured library format and wide-area network access mechanism that facilitates data discovery and exchange among GIS users at more than 80 offices of the U.S. Geological Survey. The DSDL framework includes a data dictionary of available data layers and their attributes, a standard data storage directory structure and file nomenclature, an index of available data at various map scales, data browsing software, a repository of data processing software, on-line data automation documents, and -- as they are made available -- populated spatial data layers at scales appropriate for hydrologic analysis. A workplan is being used to define priority data layers, data automation procedures, and an estimate of resources required to populate portions of the DSDL. Population of the DSDL with data, programs, and data automation guidelines is accomplished through national programs and individual offices in a coordinated manner to support existing research and water-resources applications at the local, regional, and national level.

INTRODUCTION

The Water Resources Division (WRD) of the U.S. Geological Survey has been actively applying GIS technology to hydrologic problems since 1984. In most project applications of GIS, a set of spatial data files or "layers" is created for a specific study area to prepare input to a hydrologic simulation model, to automate base map materials for field survey or site evaluation, or to perform quality assurance checks on water data collection sites in their environmental context. Traditionally, when a project is completed, the results are presented in one or more formal reports, and the data used in the investigation are stored on tape or an archivable medium in the format of the investigator's choosing. Unfortunately, much of the information developed in an investigation -- a large portion of it being spatially referenced -- is not readily available to others for potential re-use or evaluation. As a result, some spatial information must be re-digitized or processed in subsequent studies, leading to a duplication of effort.

In addition, the techniques used to automate the spatial data are not often well documented. This may lead to differences between the resulting data sets produced from the same original source material. Different assumptions used in the processing of the spatial information could affect its utility in a given application. At present, the user of spatial data has to assume that spatial data to be used in a project is of adequate quality and resolution for the hydrologic task at hand.

The National Water Information System (NWIS), the data management system used by WRD to store, process, manage, and disseminate water-related data, is undergoing a complete redesign in order to integrate what had previously been separate discipline-oriented water resource data bases into a single, relational structure (U.S. Geological Survey, 1991). User groups representing the various disciplines have identified requirements for spatial data query functions using both NWIS site information and external or "reference" spatial data. Although the spatial query functions have been identified by various user groups developing the NWIS design, the structure and existence of these reference spatial data sets is deemed to be outside the scope of the NWIS design effort.

New emphasis has been placed on the value of digital information within the federal government, and on digital spatial data in particular. Office of Management and Budget (OMB) Circular A-130 specifies that agencies must make information available in digital form and make more information electronically-accessible through on-line computer networks such as the Internet or through dial-in bulletin board services. The Federal Geographic Data Committee, also authorized by OMB, began a prototype spatial data clearinghouse activity in late 1992 to identify spatial data holdings, their fitness for use, and availability. Proposals to develop a National Information Infrastructure and a National Spatial Data Infrastructure underscore this general requirement to organize and document spatial data holdings and to improve access to them.

Statement of Problem

Geographic information systems (GIS) software does not currently provide for the management of descriptive information about data sets, known as metadata, or permit discovery and retrieval of such data, computer programs, and recognized techniques among users in either a local- area or wide-area network setting. A need exists in the Water Resources Division and the U.S. Environmental Protection Agency (USEPA) to develop an on-line system to manage digital spatial data, associated computer programs, standardized procedures and data set naming conventions, and related documentation to support a variety of hydrologic and data base applications in a distributed processing environment. The U.S. Environmental Protection Agency (USEPA) is assisting in the funded development of this on-line system with the WRD.

Purpose and Scope

The intent of this report is to present a conceptual design for an on-line library of spatial data, programs, and documentation to support hydrologic applications in the Water Resources Division of the U.S. Geological Survey. The library concept shall include definitions of 1) format and organization of spatial and attribute data, programs, and associated documentation used by Division data bases and applications programs, 2) programs required to query and evaluate the data stored in the library, and 3) the anticipated processing environment in which the library would operate. This report will be used to implement and populate the Distributed Spatial Data Library (DSDL).

Background

In February 1991, a group of geographic information system (GIS) professionals from WRD, National Mapping Division (NMD), and Geologic Division (GD) was convened to define the concept of a distributed spatial data management system that would reduce duplication of effort in data development for GIS and promote sharing of programs and consistent data across the wide area network. The need was recognized to extend information accessibility beyond what is available in the GIS continuum, electronic mail, and word of mouth.

For a prototype, the following elements were provided by the group to define its scope:

O Design and initial implementation will be in ARC/INFO[1], the predominant GIS package in use by the Water Resources Division, to expedite prototype development and integrate with the GIS processing environment, but to plan for implementing portions of the software (e.g. data posting and access) using public-domain or non-proprietary software.

O DSDL is a distributed repository with custody of specific data sets assigned or shared in an explicit manner, enforcing ownership of the data and programs.

O Documentation of data sets (known collectively as metadata) will be required for all data stored in DSDL. The DOCUMENT program, written by WRD and USEPA participants, is to be used to document spatial data officially in WRD[2].

O Programs to build, query, and manage spatial data and program holdings at a site must be written. This shall support local and remote access to the actual *information*, not just an index to holdings.

It was clear that the group wanted a system as automated as possible so as to take the burden of providing data to users and placing it on an automated information server. One participant suggested that with proper review, placing these data sets on-line through DSDL could be the mechanism to make them publicly available, not just to USGS, akin to publishing the digital data. Whereas the details of reviewing and publishing digital spatial data are being worked on, the concept of offering incentives rather than mandates to GIS data developers was well received.

An overall design was prepared and presented at several GIS conferences in 1991. This design reflected the scope of the programming and data organization that would be required. This was a conceptual design rather than a physical design. As such it is important to think of this design also in a distributed environment, inside or outside of a GIS package.

---

[1]Use of trade names does not constitute endorsement by the U.S. Geological Survey

[2]Since the initial development of the DSDL support software, the FGDC has developed a draft Spatial Metadata Standard which is anticipated to be incorporated into GIS software by vendors during 1994. Metadata management software which is compliant with the FGDC standard may eventually replace the DOCUMENT command written by the USGS and EPA.

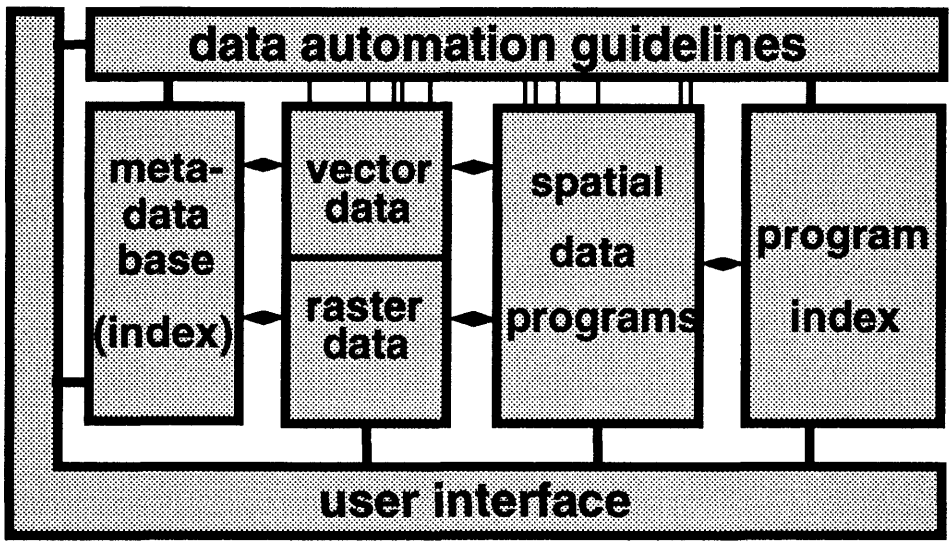

Figure 1. Conceptual organization of DSDL

Figure 1 illustrates the conceptual organization of DSDL. At the core of DSDL are collections of vector and raster spatial data sets and programs that work on specific or general types of data. The data sets and, in particular, their documentation can be indexed by textual content and by geographic location, or *footprint*. It is desirable to have a program to allow anyone to browse existing data holdings in advance of a new project to avoid reprocessing or digitizing of existing data. One should be able to query the holdings of a database based on a combination of text terms and geographic extent, through the interactive definitions of a polygon on the screen, and have potential matches listed and accessible.

Much duplication of effort has occurred writing the same GIS utilities from scratch in different offices. Whether these programs are in standard use (a formalized procedure) or are just handy, they should be able to be shared or posted. A convenient way to organize the programs and make them available is needed. Likewise, an index to the programs that is amenable to random text query is needed to enable the discovery of suitable programs across the wide area network.

A more organizational element of DSDL is the development of accepted data automation procedures, written up as guidelines. These would formalize, and hopefully standardize, methods of data acquisition or processing. Producers could reference the appropriate method and users could expect a known level of quality in the information product. End-users following such guidelines would produce data sets that could be combined into regional or national data sets for display or analysis. Multiple lines are shown between the guidelines and data and programs in figure 1 to illustrate the multiple possible cross-references that could be stored in the system.

Wrapped around all of this is a user interface, encompassing the many programs the user or administrator would run that facilitate the use of the system. Functions include programs that facilitate data administration, data query, data display, and data retrieval. The more readily available the information is, the more likely it will be used and re-used.

The basis of DSDL -- applying a distributed data base approach to disparate collections of spatial data -- was adopted by the Federal Geographic Data Committee (FGDC) in the summer of 1993 as part of a spatial data clearinghouse prototype. The clearinghouse prototype involved more than 60 participating federal, state, and private organizations in the preparation and on-line dissemination of descriptions of spatial data holdings. The U.S. Environmental Agency has also begun to apply DSDL concepts and software in the development of its spatial data management plan and is co-developing the software tools required to make DSDL a reality.

PHYSICAL DESIGN

The DSDL programs require a basic directory structure for data and documentation that is based on the ARC/INFO Librarian data directory structures but is enhanced to incorporate raster and vector data in the same reference framework. It is critical to the success and acceptance of a library concept by the end user, that all data types are identifiable and accessible through a common set of tools. The concepts of a grid collection, an image catalog, and a vector data library are all supported by ARC/INFO GIS software but the information contained therein is scattered across the system and is accessible by a wide range of commands. the structures described in this section build upon the vector library structure as a basis for referencing spatial data sets. This directory structure is illustrated in figure 2.

Data Library Framework

The map library structure is referenced by a system-wide file in the INFO relational database manager that can be read by all users of the GIS software at the local site. This file, known as the library locator file, sets an alias between the common name of the library, assigned by the database administrator, and a disk location of the library structure directory. The library structure directory includes subdirectories that include templates of each entered data layer, data set access, and a polygon data set which defines the extent of one or more geographic tiles or partitions to the data base and the location of each tile directory on disk. This hierarchy is diagrammed in figure 2, modeled after the structure of a library in ARC/INFO (Environmental Systems Research Institute, 1991).

Because all libraries are referenced in the top-level locator file, programs can access this information and create, update, or read required information from any layer associated with any library on the system without the user needing to know the actual disk location. In fact, by using existing programs or the DSDL enhanced programs, data may be moved around on disk or shared between computer systems without the knowledge of the end-user. This is an advantage to the end user who would otherwise need to remember long pathnames that might be subject to change. It was the stability of the library locator file and the predictable data structures that have been used to extend the library data model to include raster and image data collections, despite the fact they are not directly supported by the GIS software.

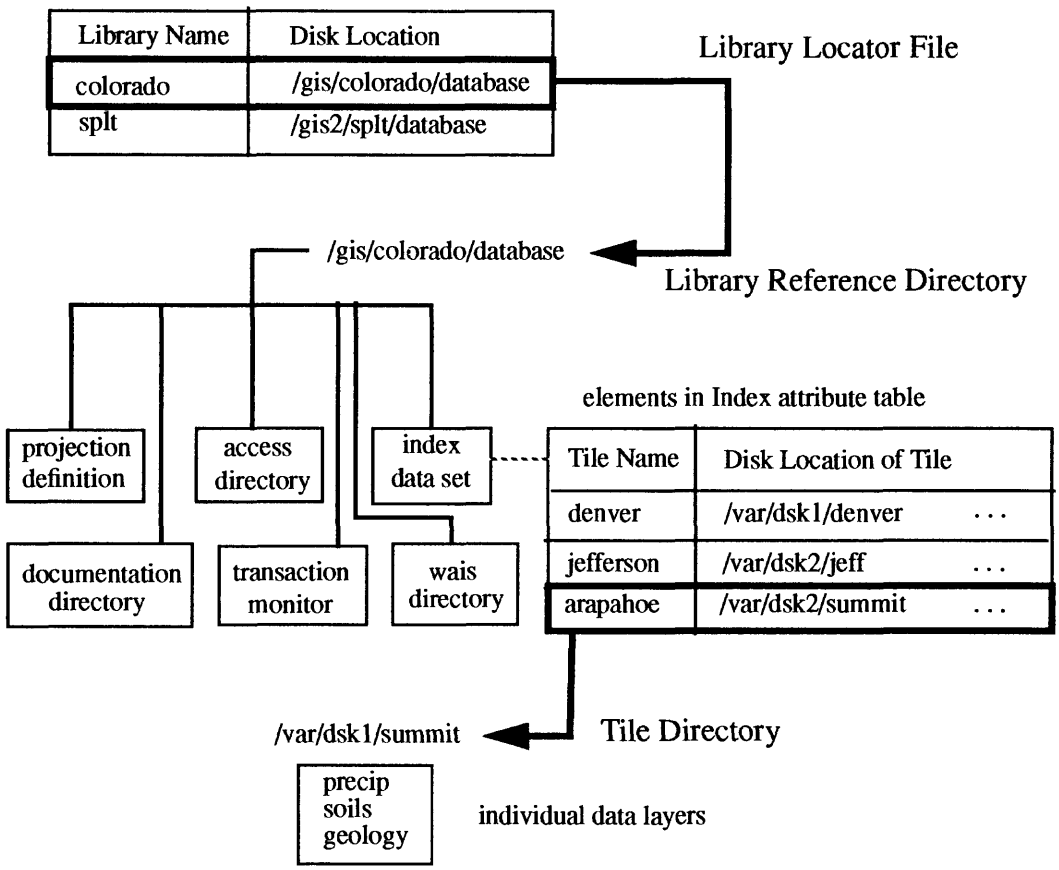

Figure 2. -- Illustration of data library directory structure.

Considerations in building a data library

A Library is defined as a collection of map themes, or layers, that have the same inherent tiling or partitioning structure. There may be multiple map libraries created and managed at a site, corresponding to different collections of data. Tiles are defined as closed geographic areas or units of map management by which data are produced, edited, and distributed. Some states build map libraries with the tiling unit being a township because that is how lands are administered. Updates are typically processed by township, reports are typically written by township, and queries are typically run against a given township. Some states manage their map data by quadrangle because it is collected by quadrangle. For scientific applications it is preferable that data be managed in units that make sense to the way the data were collected and the way it will be used. Stream traces, like the river reach file, are logically organized by river basin -- not by quadrangle. Other data may be organized by state or county or by project area.

At ARC/INFO Version 7.0 and later there is a new concept in spatial data management called Arc Storage Manager (ArcStorm). The basic premise is that data are to be managed as complete units with the partitioning or tiles being set by the software and made 'invisible' to data processing and storage. The actual data structure that ArcStorm will manage is modeled after the current Librarian structure but will be more highly automated. Another feature of the ArcStorm design is that data will be managed by theme and feature, with multi-user security supported similar to transaction processing in RDBMS. The Librarian and ArcStorm packages will work together. therefore, the DSDL programs written against the Librarian structures can be easily modified to transition into the ArcStorm structures for continued functionality.

The following are recommendations for data partitioning within the proposed DSDL structure:

Data completeness at a given scale -- If a data theme is available for the whole extent of interest and is not prone to frequent editing, then it is desirable to manage the data set as a single unit (tile) for the whole area. If the data set is incomplete for the area of interest or is prone to edits and updates by quadrangle, for example, then a library with quadrangle tiles is probably warranted.

Large-scale data (1;24,000-scale): due to its incomplete coverage in most parts of the country, it is advisable to create a 1:24,000-scale library with a tile for each quadrangle. As data are inserted into the library, queries can be posed and the data can be used as if the data were seamless.

Intermediate-scale data (1:100,000- to 1:250,000-scale): The data themes of hydrography (from the river reach file) and transportation are complete for the conterminous U.S. Hydrography data should be managed by hydrologic cataloging unit in the Library to ensure update compatibility with EPA and to encourage network analysis. Transportation data should be managed within a 1:100,000-scale tiling structure due to the size of the data sets. Land use land cover data at 1:250,000-scale has significant temporal discontinuities and therefore is not easily or appropriately map-joined. It is recommended that the 1:250,000 and 1:100,000-scale land use data be processed and placed into a quadrangle-based tiling scheme.

Small-scale data or continuous data: Point data sets and data sets that can be managed as single coverages or large extent should be managed as a single piece -- a single tile for the state or country. Data sets in this realm include population, site coverages, county and political boundaries, National Atlas data sets and 1:2-million source-scale data sets (Ecoregion, Stream, HUC2M, Land Use, NAWQA, and others). If scale is known and constant for the theme, then the layer name should reflect the intended scale of use. If the data are of mixed scale then a scale suffix is not required but should be characterized by an attribute of each feature.

Recommended library definitions

The following definitions are provided as a guide to naming libraries within a WRD office. The prefix "LIB" should be replaced with the state name or postal two-letter code, or project code in the case of a NAWQA study unit. For example, a 1:24,000-scale source collection managed in 7-1/2 minute quadrangles in Colorado should be called "co24k."

LIB24K: 7-1/2 minute quadrangle tiling structure
Example data sets:
Hydrography
Transportation
Political boundaries
Soils
Geology
Hydrogeologic units
Quadrangle index

LIB100K: 1:100,000-scale quadrangle tiling structure
Example data sets:
Transportation
Hypsography (contours)
Land Use (where collected at 1:100K)
Geology
Quadrangle index

LIB250K: 1:250,000-scale quadrangle tiling structure
Example data sets:
Land use land cover (GIRAS)
Elevation point data (DEM)
Quadrangle index

CATUNITS: Hydrologic cataloging units (8-digit code)
Example data sets:
EPA River Reach
Subbasins

STATE: (use the actual short state name or postal code) one-tile library that is defined by the state outline
Example data sets:
US population 1980, 1990
Sites (precipitation, SW, GW, WQ, etc.)
Ecoregion
Counties
Stream
Federal Lands
Public Land Survey
Geology
Hydrologic units

CUSA: Conterminous United States

In addition, a library of mapjoined and clipped data may be created for a project or study area. For NAWQA study units these should be named after its four-letter abbreviation. If you are not familiar with the code, contact the NAWQA program office to obtain it. The disadvantage to making a copy of the database for a study area is that data are unnecessarily duplicated, creating a potential update "anomaly" between the different on-line data sets -- which one is current? It should be satisfactory to restate the mapextent for your study unit and simply use and view data from the other libraries, rather than duplicating it.

The library for the conterminous United States (CUSA) is maintained at the Reston node and contains small- and intermediate-scale data sets with full coverage of the Lower 48 states. These data are prepared for state-wide, regional, and national analysis and are often used as backdrop maps or for the preparation of base map materials. The preparation of the CUSA library on CD-ROM is planned for internal distribution in 1994 for reference by the NWIS-II software.

Before creating a Library one must consider the projection system, units, tiling approach, and whether single or double precision coordinates are warranted. From these considerations you will need to create an INDEX coverage whose polygons represent the tiles.

In considering the selection of a projection system for your project/District to use, consider the following advantages and disadvantages:

Cylindrical projections (UTM, transverse Mercator)
Preserve direction. Distortion increases away from line of tangency
Multiple zones to represent large longitudinal extent (UTM)
Not good for study areas spanning more than one Zone (UTM)
Good for 7-1/2 minute maps used by themselves
Appropriate for smaller areas (300 miles wide or less)

Conical projections (Albers equal-area, Lambert conformal conic)
Preserves area and length, sacrifices direction
May have one or two "standard" parallels of negligible error
Better for data sets of large extent
Appropriate for state, national, and continental depictions with low error

The NAWQA study units are being advised to use the Albers Equal Area projection with the same parameters as the 1:2,000,000-scale DLG data, with a Central Meridian determined so that the unit is centered on a correct north-south line. Changing the Central Meridian of this projection merely rotates the map, with no other distortion in length or area introduced.

The tiling systems suggested in the previous examples should be considered in the development of libraries in a given state. For the time being it is recommended that standard libraries (as listed) be used as they are named in this document.

The names of layers to be placed in libraries should be standardized. Standard theme names and attribute names are being developed with the assistance of NAWQA project personnel and are defined in published spatial data automation guidelines released as Open-File Reports. Different procedures and names may be developed for similar data at different scales or from different sources.

Data layers/coverages of a fixed scale source: Theme name + source scale denominator (e.g. SOIL24 for 1:24,000-scale soils)

Data layers/coverages of mixed scale source: Theme name (e.g. TIGER -- from 1:24,000 to 1:250,000-scale source)

Computer Program Library Framework

The second component of the DSDL concept is the maintenance of and providing access to computer programs developed within the GIS for local use and for discovery by other WRD scientists. The ARC/INFO software model allows for the definition of computer programs in a scripting language known as Arc Macro Language (AML) and those written in compiled languages such as C and Fortran and in interpreted languages such as the Bourne shell or Perl. The AML interpreter is active at all times during an ARC/INFO session and references specific locations to find and execute Arc commands, AML programs, or other support programs written in other languages.

Programs that are written in AML and are to be executed as if they were installed standard GIS commands are located in a subdirectory under the site-specific UNIX environment variable $ARCHOME known as "atool"[3]. User-developed script and compiled executable programs are to be installed in the $ARCHOME/utool subdirectory. Subdirectories under both atool and utool further organize which programs can be executed as commands in each of the subsystems of ARC/INFO -- Arcplot, Arcedit, Grid, and others. The organization of software directories within the ARC/INFO environment are depicted in figure 3. Commands that are to be shared with other users at a site shall be placed in the appropriate atool or utool directory. A corresponding helpfile (text) is to be placed in the relevant $ARCHOME/help subdirectory. With version 6.1.1 of the software, helpfiles with illustrations may also be prepared in Frame Maker and can be accessed through the HELPVIEW command using the Frame viewer product provided with the ARC/INFO product. Software developers are encouraged to create illustrated and formatted help documents with their software to be shared with other users in the Division.

[3]At most sites in WRD, the ARC/INFO software is loaded into the /usr/opt/esri directory and the ARCHOME variable is set to /usr/opt/esri/arcexe61 (for version 6.1 of the software). The atool directory at these sites would then be either $ARCHOME/atool or /usr/opt/esri/arcexe61/atool.

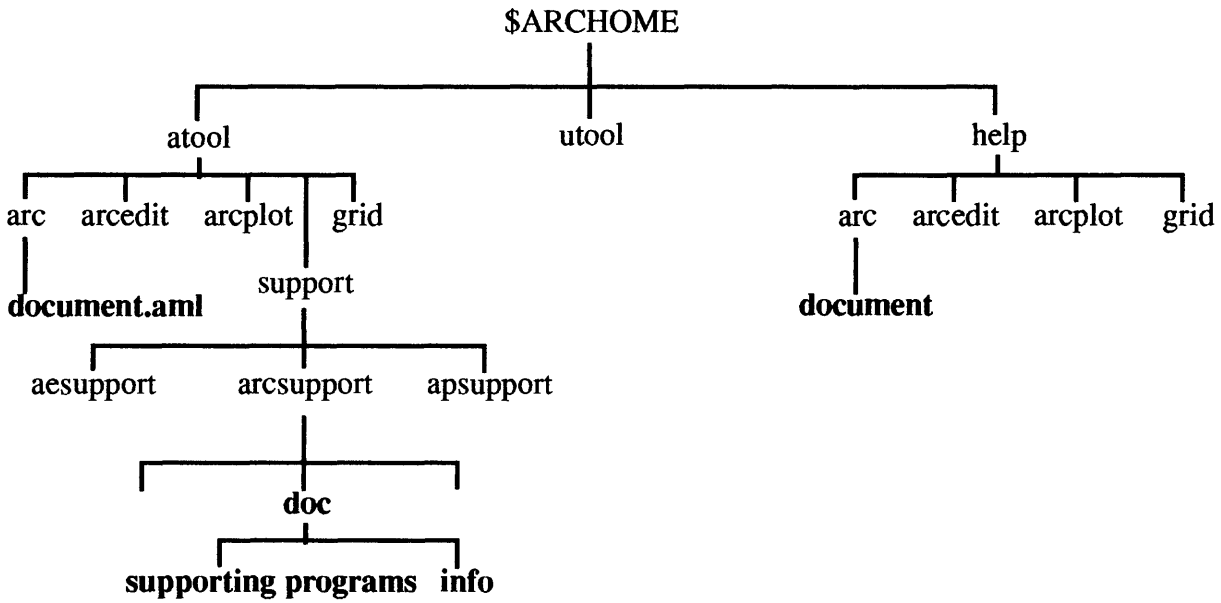

Figure 3. Illustration of program directory organization for the DOCUMENT program example

Figure 3 also shows the location of the support subdirectory under atool. Supporting programs, data base templates, menus, source code, and other files should be stored in a special directory created by the author under the appropriate support subdirectory. This allows for programs that work together to be managed in the same directory rather than be merged in with other unrelated programs, simplifying administration.

Preparing a self-installing "tar" file

Programs and supporting files, including a help file, are to be exchanged within the WRD through use of a self-installing "tar" formatted file. A UNIX tar file (meaning tape archive) is designed to encode and extract full directory structures with either explicit or implicit pathnames. Because the $ARCHOME directory may be in different locations on different computer systems in WRD, tar files should be prepared while the user is in the $ARCHOME directory under which all the programs and supporting files are located. The steps are illustrated for the DOCUMENT AML, its help file, and supporting files.

Step 1. Change directory to the $ARCHOME directory.

```
% cd $ARCHOME
% pwd
/usr/opt/esri/arcexe61
```

Step 2. Prepare a script which will load all the supporting files and directories into a named tar file and then compress the tar file. Note that the pathnames referenced in the script are relative to the current directory, $ARCHOME. the backslashes (\) are continuation characters that allow all the objects to be loaded into the tar file as if they were all entered on the command line.

```
% vi document_14.sh

tar cvof document_14.tar \
atool/arc/document.aml \
help/arc/document \
atool/support/arcsupport/doc
compress document_14.tar
:wq
```
[4]

It is important that only underscores -- "_" -- be used in the script, tar, and description file names so that the suffixes, delimited by a period can be managed by the software. It is recommended that the version number of the software being provided be included in the files being managed, although the actual command (e.g. DOCUMENT) will not include the version number in its name.

Step 3. Make the script executable.

```
% chmod +x document_14.sh
```

Step 4. Run the script.

```
% document_14.sh
a atool/arc/document.aml 89 tape blocks
a help/arc/document 32 tape blocks
a atool/support/arcsupport/doc/att.menu 3 tape blocks
a atool/support/arcsupport/doc/atttable.menu 2 tape blocks
a atool/support/arcsupport/doc/doc.menu 4 tape blocks
a atool/support/arcsupport/doc/doccite.menu 2 tape blocks
a atool/support/arcsupport/doc/info/arcdr9 6 tape blocks
a atool/support/arcsupport/doc/info/arc000nit 9 tape blocks
a atool/support/arcsupport/doc/info/arc000dat 0 tape blocks
a atool/support/arcsupport/doc/info/arcnsp 0 tape blocks
a atool/support/arcsupport/doc/info/arc001nit 1 tape blocks
a atool/support/arcsupport/doc/info/arc001dat 0 tape blocks
a atool/support/arcsupport/doc/info/arc002nit 4 tape blocks
a atool/support/arcsupport/doc/info/arc002dat 0 tape blocks
a atool/support/arcsupport/doc/info/arc003nit 1 tape blocks
a atool/support/arcsupport/doc/info/arc003dat 11 tape blocks
a atool/support/arcsupport/doc/info/arc006nit 2 tape blocks
a atool/support/arcsupport/doc/info/arc006dat 0 tape blocks
a atool/support/arcsupport/doc/log 1 tape blocks
```

[4]The :wq command is the function in the vi editor to write and quit. It is not a part of the program being written.

```
a atool/support/arcsupport/doc/acurs.aml 4 tape blocks
a atool/support/arcsupport/doc/catt.menu 3 tape blocks
a atool/support/arcsupport/doc/cattdefs.menu 1 tape blocks
a atool/support/arcsupport/doc/catttable.menu 2 tape blocks
a atool/support/arcsupport/doc/cdoc.menu 5 tape blocks
a atool/support/arcsupport/doc/check_off.icon 1 tape blocks
a atool/support/arcsupport/doc/check_on.icon 1 tape blocks
a atool/support/arcsupport/doc/datt.menu 3 tape blocks
a atool/support/arcsupport/doc/ddoc.menu 4 tape blocks
a atool/support/arcsupport/doc/dmain.menu 4 tape blocks
a atool/support/arcsupport/doc/dmainmenu.help 2 tape blocks
a atool/support/arcsupport/doc/dref.menu 2 tape blocks
a atool/support/arcsupport/doc/editor.menu 1 tape blocks
a atool/support/arcsupport/doc/switch.help 1 tape blocks
a atool/support/arcsupport/doc/switch.menu 3 tape blocks
a atool/support/arcsupport/doc/uatt.menu 5 tape blocks
a atool/support/arcsupport/doc/uattdefs.menu 1 tape blocks
a atool/support/arcsupport/doc/udoc.menu 6 tape blocks
a atool/support/arcsupport/doc/umain.menu 7 tape blocks
a atool/support/arcsupport/doc/umainmenu.help 3 tape blocks
a atool/support/arcsupport/doc/update.menu 2 tape blocks
a atool/support/arcsupport/doc/uref.menu 3 tape blocks
```

Step 5. Move the text explanation of the program (ending in .text), the tar file creation script (ending in .sh), and the compressed tar file (ending in .tar.Z) to the subdirectory of the $ARCHOME/software directory for which the program is valid.

```
% cp document_14.sh $ARCHOME/software/arc
% cp document_14.tar.Z $ARCHOME/software/arc
% cp document_14.text $ARCHOME/software/arc
```

Once the software is located in these directories it can be searched via a WAIS server or fetched via a network library server (NetLIB) developed by the DSDL programming staff.

Installing a self-installing "tar" file

Once a tar file with programs is retrieved it must be uncompressed and then copied to the $ARCHOME directory. the user must have adequate permissions in this directory to install new software because new directories may be created and old files overwritten (in the case of updates). The tar command is used, once again, but using its extract (x) option.

```
% cd $ARCHOME
% ls doc*
document_14.tar.Z
% uncompress document_14.tar
% ls doc*
```

```
document_14.tar
% tar xvf document_14.tar
x atool/arc/document.aml 89 tape blocks
x help/arc/document 32 tape blocks
x atool/support/arcsupport/doc/att.menu 3 tape blocks
x atool/support/arcsupport/doc/atttable.menu 2 tape blocks
x atool/support/arcsupport/doc/doc.menu 4 tape blocks
x atool/support/arcsupport/doc/doccite.menu 2 tape blocks
x atool/support/arcsupport/doc/info/arcdr9 6 tape blocks
x atool/support/arcsupport/doc/info/arc000nit 9 tape blocks
x atool/support/arcsupport/doc/info/arc000dat 0 tape blocks
x atool/support/arcsupport/doc/info/arcnsp 0 tape blocks
x atool/support/arcsupport/doc/info/arc001nit 1 tape blocks
x atool/support/arcsupport/doc/info/arc001dat 0 tape blocks
x atool/support/arcsupport/doc/info/arc002nit 4 tape blocks
x atool/support/arcsupport/doc/info/arc002dat 0 tape blocks
x atool/support/arcsupport/doc/info/arc003nit 1 tape blocks
x atool/support/arcsupport/doc/info/arc003dat 11 tape blocks
x atool/support/arcsupport/doc/info/arc006nit 2 tape blocks
x atool/support/arcsupport/doc/info/arc006dat 0 tape blocks
x atool/support/arcsupport/doc/log 1 tape blocks
x atool/support/arcsupport/doc/acurs.aml 4 tape blocks
x atool/support/arcsupport/doc/catt.menu 3 tape blocks
x atool/support/arcsupport/doc/cattdefs.menu 1 tape blocks
x atool/support/arcsupport/doc/catttable.menu 2 tape blocks
x atool/support/arcsupport/doc/cdoc.menu 5 tape blocks
x atool/support/arcsupport/doc/check_off.icon 1 tape blocks
x atool/support/arcsupport/doc/check_on.icon 1 tape blocks
x atool/support/arcsupport/doc/datt.menu 3 tape blocks
x atool/support/arcsupport/doc/ddoc.menu 4 tape blocks
x atool/support/arcsupport/doc/dmain.menu 4 tape blocks
x atool/support/arcsupport/doc/dmainmenu.help 2 tape blocks
x atool/support/arcsupport/doc/dref.menu 2 tape blocks
x atool/support/arcsupport/doc/editor.menu 1 tape blocks
x atool/support/arcsupport/doc/switch.help 1 tape blocks
x atool/support/arcsupport/doc/switch.menu 3 tape blocks
x atool/support/arcsupport/doc/uatt.menu 5 tape blocks
x atool/support/arcsupport/doc/uattdefs.menu 1 tape blocks
x atool/support/arcsupport/doc/udoc.menu 6 tape blocks
x atool/support/arcsupport/doc/umain.menu 7 tape blocks
x atool/support/arcsupport/doc/umainmenu.help 3 tape blocks
x atool/support/arcsupport/doc/update.menu 2 tape blocks
x atool/support/arcsupport/doc/uref.menu 3 tape blocks
```

Guidelines and Document Framework

The final framework element of the DSDL implementation provides for the management of written documents describing techniques applied in data base development, findings of investigations, and project summaries. Most important of these are documents that describe, as mentioned in the conceptual diagram in figure 1, spatial data automation guidelines. Spatial data automation guidelines are instructions on how to prepare a standardized data set from a specific medium -- paper or digital.

Standardization of data sets allows data bases such as NWIS-II and organizations such as the National Water Quality Assessment Program build consistent references between applications and data entities. Through the publication of these guidelines and associated programs, the reproducibility of data sets are ensured, resulting in a form of certification of the data sets. Data produced to these published guidelines may reference the guideline rather than repeat the techniques used.

Guidelines and other documents are to be stored (or have symbolic links set from) the $ARCHOME/documents directory. Files which are ASCII text files shall be suffixed with .text, whereas FrameMaker documents are to be suffixed with .doc and PostScript files are to be suffixed with .ps. These file suffixes are designed to assist in data discovery and presentation.

FUNCTIONAL SPECIFICATIONS

Based on an evaluation of functionality supplied within common relational database management systems and geographic information systems software, a set of core functions was identified by the design team as the basis for a viable spatial data management system that operates both inside and outside of the ARC/INFO software. These functions are explained as specific work elements which are roughly equivalent to a command-level program. Programming support for most of these functions was completed and made ready for release in December 1993. All commands listed here are supplied with on-line help files within the ARC/INFO system.

Data Administration Functions

The following elements are considered functions required for the administration of the DSDL software and libraries. These functions are to be executed by a spatial data base administrator (DBA) with specific privileges at a site.

1. Create a library structure for data sets. A program was created in Arc Macro Language (AML) to consolidate the functions of Arc Librarian into a single procedure. This program is called CREATELIBRARY and is executed within the GIS environment. The program creates a librarian directory structure with support for data documentation.

2. Insert a data layer into library. A program was created to insert spatial data layer into a library while verifying the presence of spatial data documentation files. The program is called INSERTLAYER.

3. Install a program to a common location. This function is described in the Computer Program Library Framework section of this report.

4. Post a program and its documentation for query and retrieval. Programs, description files, and self-installing tar files can be served, searched, and retrieved using DSDL client-server software including the Wide-Area Information Server (WAIS) software.

5. Update a data set. Programs already exist within Librarian to perform a data set update. A program has been written to simplify the data update process in a transactional environment (UPDATELAYER).

6. Change physical location of data set. Program was written (MOVETILE) to enable the DBA to change the physical location of a tile and the spatial data sets it contains from one directory to another and to pull the data set off-line while maintaining documentation. If the tile is taken off-line, a reference as to how to obtain the off-line data is stored in the active library.

7. Remove data set from Library. Program was written to remove the contents of a spatial data set from all, or a portion, of the full data extent. One partition (tile) worth of data may be removed or an entire layer (with documentation) may be removed, at the DBA's discretion. May be invoked by DBA only. Program is called DROPLAYER.

8. Set user access to portions of Library. A program has been written to assist DBA in granting and revoking various data access privileges. The program allows for the ability to insert coverages, images, software, and guidelines and incorporate GIS and UNIX file-level access. Program is called SETACCESS, and may be invoked by DBA only.

9. Create and update central data dictionary and guideline repository. This is not a generalized DSDL function but a centralized one under development through a WAIS server. A data dictionary template is being developed in concert with NWIS. Into this template will be placed all NWIS-II data dictionary elements as well as documented elements that are members of nationally accessible data sets. Field users may also forward data elements electronically to be incorporated into the server by DSDL staff. This function will be included in a future release of the software.

10. Develop a local data dictionary from themes stored in all libraries. A program has been written to extract all attribute definition information at a local site and place it in an on-line format identical to that described in Task 9, above. This data dictionary can be accessed locally or remotely using the WAIS software. This function will be included in a future release of the software.

11. Extend the data model support to include raster data structures. Data documentation (DOCUMENT) and all library functions shall be available for all data types including vector, grid, and possibly image data. This information is required for the presentation of all available information via BROWSELIB. This function will be included in a future release of the software.

12. Serve subset of local Library contents to wide-area network. Support programs have been written to convert available GIS data in vector and raster format to formats suitable to WAIS. There shall be three format files for each data set to be served -- one for the formatted metadata, one (or potentially multiple) graphic 'snapshots' in graphics interchange format (GIF), and one for the exported data in compressed format. The program is called ARCWAIS and includes a menu interface to assist the DBA in the population and serving of the spatial data. spatial data file shall also include a geographic outline or 'footprint' that can be used by the spatial WAIS server in order to search and retrieve based on both text and spatial criteria.

13. Track retrieval of DSDL data by remote users. A subscription program has been written in perl programming language to evaluate access of any given WAIS server by user. DBA may then evaluate the type of access and inform users by electronic mail as the content of data or documentation changes. The program is called SUBSCRIBE. This function will be included in a future release of the software.

14. Menu-driven interface to spatial data administration functions. A menu system has been written to help a spatial DBA administer the DSDL services at a site. The ability to search for word constructs and commands shall be included in this composite form menu modeled after HELPMENU. This function will be included in a future release of the software.

Browsing and Retrieval Functions

1. Invoke a program from common location. This feature is already accommodated by the ARC/INFO GIS software. Programmers that create self-installing programs and helpfiles will provide for these programs to be discovered and invoked from the appropriate portion of the GIS software.

2. List available commands (programs) and syntax. This feature is already provided for by the ARC/INFO package using the COMMANDS and HELP programs provided that the help files are properly installed under $ARCHOME.

3. Browse available local data sets. Two programs have been written to accomplish this task. BROWSELIB allows a user to evaluate the contents and documentation of all layers in all libraries on a given local data system through a use of linked scrolling menus. LIBRARIES is also available to display the available layer names in all libraries.

4. Display all data types from Library. Vector and raster data sets are supported by the GIS software, but the Librarian package does not provide for anything other than vector data. This enhancement is being provided under data administration task 11, above. The program SETLIB has been written to provide access to common grids and images in a library from an ARCPLOT session.

5. Extract a data set from Library to current directory. EXTRACT program has been written to clip out a selected layer from a library based on a defined closed polygon data set in the same or different projection.

6. Browse remote library contents and retrieve documentation and data. This functionality is provided with the DSDL-enhanced WAIS software package and the network library access package NetLIB under development by the DSDL staff.

7. Display data from Library using symbology and generalization rules. Drawing instruction files may be included with a library layer to produce selective symbolization of features present. This functionality shall be invoked by using the DRAWCOV or DRAWGRID command. Drawing instruction files shall be placed in the library in a commonly accessible location. The DRAWCOV programs shall be placed within a coverage directory for a single coverage or in the database directory (under Arc 6.1.1) as AML programs. Different types of data presentation may result in different DRAWCOV programs being written. Base map generation, workstation display, and electrostatic plot outputs all require different color sets, text fonts, and settings and will be developed by the authors of the data sets as the need arises. This function will be included in a future release of the software.

8. Retrieve data elements or data automation guidelines by search keywords and display or print results. This functionality is provided with the standard WAIS software package.

DSDL Software

Software to facilitate the implementation of a DSDL library is included on a diskette with this report. The diskette includes all programs, helpfiles, and instructions to install the DSDL support software. The DSDL support software is designed to simplify the creation and maintenance of on-line documentation and spatial data libraries for a variety of data types supported within the ARC/INFO geographic information system (GIS) software. The initial release is intended for the Data General Aviion UNIX workstations operating at DG/UX 5.4.2 or higher. To conserve space, only executables are included in this release. Source code and instructions for compilation on other platforms can be obtained from the authors.

Each command on this 1.4MB 3-1/2" diskette is represented by a pair of ASCII text files to assist in understanding thesoftware. These include the program documentation comments (.hdr), and a help file (.hlp). The entire suite of software, designed to be installed in ARC/INFO under UNIX is included in the file dsdlamls.tgz -- a UNIX tape archive file (tar) which has been compressed using the GNU ZIP utility. The file INSTALL.TXT provides step-by-step installation instructions for the suite of software from the .tgz file. The file (README.TXT) provides a general overview of the software.

Interaction with GIS Software

The design group identified an overall requirement that the software be compatible with the predominant geographic information system software used by the WRD. In April 1993, a department-wide contract was let to Environmental Systems Research Institute for the ARC/INFO suite of GIS software. As a result of this seven-year contract, prototyping of the DSDL programs done with the ARC/INFO software could continue without the need to convert the programs to a different GIS environment.

There was also a recognition that a portion of the interface to our collections of digital data not rely on proprietary software. In this way, access to approved digital spatial data sets could be provided to outside users potentially using other brands of GIS or database software. One requirement of the GIS contract states that translators must be provided by the GIS vendor to comply with the Spatial Data Transfer Standard (SDTS), Federal Information Processing Standard 173, soon after specific profiles (subsets of the standard) are published by the National Institute for Standards and Technology (NIST). Compliance with this standard will allow users of the GIS and the DSDL extensions to provide data in non-proprietary formats for the first time without loss of information.

National Water Information System interface

The National Water Information System, version II (NWIS-II) is an integrated data base to manage all types of water-related information. The evaluation of water-related data in a geographic context was identified as a requirement for the NWIS-II software by the NWIS-II user groups. Such functionality will be supported in a future release of the NWIS-II software and will provide access to common-use spatial data in a DSDL library as well as data sets developed by the end-user. Reference to data libraries with a common definition throughout the WRD will make it possible to interface NWIS-II to DSDL structures.

DATA REQUIREMENTS

In the Spring of 1989, each of the NWIS User Groups was solicited for input on what "standard" spatial data layers were most important to their discipline. In addition, information was requested as to the appropriate scale and coverage required for investigations. The results are given in Table 1.

The survey appears to reflect an awareness of the spatial data at 1:2 million scale that was created by National Mapping Division but processed and made available to WRD users by the National Water Summary. Although the scale is coarse, most states have acquired these data sets for demonstration projects and for the compilation of small-scale base maps and illustrations. Most user groups indicated that the layers should be prepared from 1:100,000-scale materials and ultimately would like to see 1:24,000-scale coverage as it becomes available.

The mixed ranking of the Public Land Survey System (PLSS) item reflects the regional importance of this data base in hydrologic site locations. Whereas most western states locate wells relative to the PLSS, few eastern states have such a system or need. The mixed ranking for the TIGER files is probably due to an unfamiliarity with this relatively new data product which includes a selected subset of hydrography, roads, and census divisions at a nominal 1:100,000-scale.

Although the survey did not indicate the availability of each theme at a given scale, several groups encouraged the use of existing data prepared by other agencies such as EPA and the Census Bureau. It is the goal of this library design effort to use existing data where possible to actually populate the library. Where data are not widely available -- soils information, for example -- this effort would define a "template" format for the theme to include a minimum set of spatial features (points, lines, or areas) and an attribute set definition to include a minimum set that will be included in every data set in the Library within WRD and an extended set of attributes whose population -- but not definition -- is optional. In this way, the standardization of spatial data can be encouraged with a minimum effort on the part of the individual District.

Spatial Data Theme	SD	QW	SW	WU	SP	Average
Surface Water Hydrography	1	-	1	1	1	1.0
Hydrologic Units	1	-	3	1	1	1.5
State/County	1	-	4	1	1	1.75
Elevation	1	-	1	-	5	2.3
Geologic Maps	2	1	5	-	5	3.25
Soils Maps	2	2	5	-	5	3.5
Land Use/Land Cover	2	4	6	-	4	4
Transportation	2	5	6	-	3	4
Public Land Survey	6	3	6	-	2	4.25
Census TIGER	3	6	6	-	4	4.75

Table 1. NWIS-II user group spatial data requirements, 1989. Numbers in each column represent the importance of a data theme to a given discipline, where 1 is highest importance, and 6 is lowest. The two-letter abbreviations represent the five user groups that responded to the ranking information request (SD=Sediment, QW=Water Quality, SW=Surface Water, WU=Water Use, SP=Spatial, GW=Ground Water).

Available Data Layers

At 1:100,000- and 1:24,000-scale, the scales referenced by most end users, few data sets exist to provide full coverage for hydrologic applications that occur in study areas whose size ranges from tens to hundreds of square miles. Increasingly, projects such as the National Water Quality Assessment program (NAWQA) require basin-wide data coverage at intermediate scales (approximately 1:100,000- to 1:250,000-scale) -- coverage that often spans state boundaries.

The DSDL project has acquired, processed, and documented a number of intermediate-scale digital spatial data sets for use by the Water Resources Division. These data, once reviewed and approved for public distribution, will be made available to WRD offices on CD-ROM media as a read-only library of ARC/INFO data sets in 1994. Portions of the data will also be made electronically accessible across the wide-area network known as the Internet using enhanced public domain WAIS software.

The following data sets are available from the Reston DSDL library for the conterminous United States:

Locations of NASQAN benchmark stations, USGS-WRD

Comprehensive Environmental Response, Compensation, and Liability (CERCLA) Sites, USEPA

Climate Divisions, National Oceanic and Atmospheric Administration, 1:7,500,000 scale

County boundaries of the conterminous United States, USGS-NMD at 1:2 million scale, Bureau of the Census at approximately 1:100,000-scale

Ecoregion Map, USEPA, 1:7,500,000-scale

Installation Restoration Program (IRP) sites, USEPA

Landfill locations, USGS-WRD

National Atmospheric Deposition Program/National Trends Network, USGS-WRD

National Water Quality Assessment (NAWQA) study units in the conterminous United States, 1:2,500,000-scale

Resource Conservation and Recovery Act (RCRA) sites, USEPA

Large reservoirs in the United States and Puerto Rico, USGS-WRD

Stream Flow Basin Characteristics File, 1986, USGS-WRD

State boundaries of the conterminous United States, USGS-NMD at 1:2,000,000 scale, USGS-WRD at 1:100,000 scale and 1:7,500,000

Hydrologic cataloging units (river basins), USGS-WRD at 1:2,500,000 scale and 1:250,000 scale

Streams from the National Atlas, USGS-NMD, 1:2,000,000

1990 Census data points on population and housing units, Bureau of Census

Additional information on these data sets is available via the WAIS server from the Reston node of DSDL.

Data Access in a Distributed Environment

Originally, it was anticipated that specialized software would need to be developed to provide non-proprietary access to spatial data, programs, and documentation such as guidelines and data dictionaries. In late 1991, the Wide Area Information Server (WAIS) software was identified by the U.S. Geological Survey as a potentially useful public-domain software system for the dissemination of earth science data. The WAIS software was jointly developed by Apple Computer, Thinking Machines Corporation, and Dow Jones and Peat Marwick as a means to rapidly index and provide access to searching through entire text documents using unstructured text queries across a wide area network. Early tests of the WAIS software included the Earth Science Data Directory and the WRD Selected Water Resources Abstracts, both of which were previously accessible only through a text-based interface on a floppy diskette. The software proved robust enough to permit custom enhancements such as field-like data retrieval, spatial indexing and retrieval of documents, and support of multiple file types (e.g. a text file, a graphic file, and a binary data file all as a result of a single query) by USGS and collaborating authors.

The WAIS software is being used by the DSDL project for the posting on-line of digital data set documentation, graphic 'snapshot' files, compressed GIS data layers for retrieval, data dictionary entries, and software available within the USGS. Spatial extensions to WAIS were done in concert with the Federal Geographic Data Committee, Geo-spatial Data Clearinghouse Workgroup to provide internet accessibility to data and descriptive information held by primarily federal users of GIS.

The concepts of DSDL and WAIS rely on the fact that the data most likely to be updated and properly maintained are those kept in the local office and are in constant local use and scrutiny. Rather than create a centralized repository of digital spatial data, individual offices with GIS capability will manage and post spatial data, programs, and documents from the local site. A central directory of servers -- a who's who of digital spatial data will be created by WRD DSDL office, but only as a referral service to aide the discovery of the actual data distributed in the field. This directory of servers will also reference other known catalogs of earth science information known to be held in WAIS servers by the WRD and other earth science agencies.

Custody of Spatial Data

Distributed management of digital spatial data sets will require the management of information relating to the ownership or update responsibility of the data in the DSDL library. Such information shall be managed using the DOCUMENT command. Where data sets cross state boundaries, and the custody is shared, the update authority will need to be worked out between the affected offices. This issue of custody may also be shared with other cooperator agencies which have had a significant responsibility for the production or initial release of the data. Coordination with these external entities is encouraged and such information must be included for transfer with the data set.

PUBLICATION OF DIGITAL SPATIAL DATA

Four types of computerized data bases are described in SM Chapter 500.24.1 based on the data content, documentation, review, and quality assurance procedures. These data base types include:

- Proprietary: contains proprietary information obtained from private sources, which must not be published or disclosed outside of the government without permission.

- Internal: designed and maintained for internal U.S. Geological Survey (USGS) use only.

- Provisional: includes documentation of the origin of the data, data collection method, peer review process, and data base description. May be released to the public upon request.

- Director Approved: must contain only data that have undergone rigorous, documented quality assurance. Must be supported by operating division through a designated data base administrator. Must exceed criteria for provisional data base. May be released to the public upon request.

Examples of proprietary data bases include copyrighted data purchased from companies that restrict the rights of duplication and distribution, including privatized satellite imagery and ZIP code boundary files. Some data are considered proprietary--such as satellite imagery only in their original form. Resampling, performing classification, or changing the projection of the imagery constitutes manipulation that is not reversible. Such processing may create a new data set that is distinct from the original and is not proprietary by the specific terms of our licensing agreement. Users of proprietary data should be aware of the licensing and usage limitations of the data and derivative products.

Examples of internal data bases include administrative data bases, project data bases, and data bases produced for the USGS by outside sources. Internal data bases are considered working data bases and may contain confidential or organizational material. The content of internal data bases may not be shared with any users outside the USGS including Federal and State cooperative agencies. Data bases that have undergone formalized review for either provisional or Director-approved data bases are accessible to the public. Documentation of the content and limitations of internal data bases is suggested as a preliminary step to making the data publicly-accessible in the future.

Provisional data bases would include most of the spatial data that are to be used with basic data collection activities, investigations, and for general-purpose access in an office. These data have been collected and prepared using accepted or published procedures. Release of such data would require full documentation and evidence of a formal review certifying conformance with referenced data collection and automation procedures. Review prior to release of digital spatial data is described under a later section in this section.

Director-approved data bases include spatial and non-spatial data stored by National Water Information System (NWIS) and NWIS-II, and in the future will include spatial data bases of national scope including hydrologic unit boundaries, study unit and administrative boundaries, and other themes. Data in this category are more rigorously reviewed and quality-assured than for the provisional data category. In addition, a data base administrator must be identified to maintain the data through its life-cycle.

Documentation of Spatial Data Sets

A minimum suite of documentation is to be maintained for all spatial data bases created or used in WRD. Complex hydrologic analysis and display involve assumptions of adequate data resolution and timeliness and that the appropriate spatial data layers are used in hydrologic analysis. The revised SM Chapter 500.24.1 also requires collection and reporting of such data. The collection of descriptive information provides background information on a data set for the end-user to evaluate its fitness for use. Documentation does not necessarily imply certification or conformance with some accuracy standard, unless documented by the author. To this end, all types of digital spatial data may be documented -- including data sets for use in illustrations that should not be used in analysis.

The WRD and EPA have developed a spatial data documentation program (DOCUMENT) that facilitates the collection and management of important metadata.[5] This program was written to collect information that will be mandatory for data transfer using the Spatial Data Transfer Standard (SDTS), a Federal Information Processing Standard (FIPS) to be put into action over the next year. The DOCUMENT program shall be used to document any spatial data set that will be used in hydrologic investigations, preparing soft-copy or hard-copy maps, or will be accessed by more than one user.

The DOCUMENT program manages four types of information--basic data set characteristics, a data dictionary, references to published source(s), and a narrative section for extended discussion of data automation techniques and revisions. Where such information exists, all four types of information shall be collected and managed for all types of spatial data bases to facilitate appropriate use and re-use of the data.

Review process for digital spatial data

Provisional and Director-approved data bases that incorporate spatial references shall be reviewed for:

Positional accuracy and precision with respect to source
Contextual accuracy
Attribute accuracy
Logical (Topological) consistency

The results of this review will be included in the spatial metadata.

Positional accuracy of data in the digital product shall be described along with the methods used by the author to determine an estimate of accuracy. A deductive estimate would refer to errors introduced at each processing step. A comparison with source material on stable-base material should also be made, where available, to verify the accuracy of the resulting product and the precision with which it was reproduced. All features from the source data set should be reproduced in the data set to be reviewed unless the data set represents a documented subset of the original. Spatial data sets that meet National Map Accuracy standards shall include in their documentation the phrase: "This digital spatial data set complies with national map accuracy standards."[6]

[5]Metadata are information about the structure or content of a data base, or "data about data."

[6]National Map Accuracy Standards (NMAS) are met when 90 percent of all points tested are found to be within 1/50th of an inch (0.05cm) of their true location on a stable-base map. This fraction equates to a minimum resolution of 40 feet at 1:24,000-scale and 167 feet at 1:100,000-scale. NMAS only apply to map series at greater than or equal to 1:250,000-scale, e.g. 1:24,000- to 1:250,000-scale.

Contextual accuracy review includes the evaluation of the spatial data with other similar spatial data for the same time, scale, and areal extent. The documented scale of intended use should be supported by this review. An example of such review would be the display of surface water data collection sites with respect to streams and contour data, captured at a similar scale. Sites not falling on streams would be suspect and should be reported by the reviewer to the author for revision. A temporal review of the data is sometimes appropriate where spatial data represent a specific time period. Review of the extent of spatial features with respect to other map data for the same time period and scale is recommended, where available. Field experience or "ground-truthing" may be required for a review of data such as land use or irrigation patterns.

Attribute accuracy shall be carefully reviewed. The tables and columns (items) described by the author shall be carefully compared with the contents of the data set, what is reported in the documentation, and what is available from source materials. Attribute domain--the list of possible values for a given attribute--shall also be checked against the values discovered for each column or attribute. Table and attribute names and pathnames should be verified. Where a source map exists, the attributes in the data set shall be verified against those displayed on the source map. Where a source map does not exist the author shall note "No published source map available" in the narrative section describing source.

Topological consistency or integrity of the spatial data set should be checked. For example, if area features are part of the data set then they should be described in the documentation. If the data set is exclusively a polygon data set and "dangling" or orphan line segments are not appropriate, they shall not be included in the data set. All area features shall have polygon labels. Point and line features should also be so documented. Raster data sets shall be documented as to the number of rows and columns and layers, if appropriate. These shall also be verified by the reviewer. Projection information shall be maintained with each data set and its presence shall be verified by the reviewer.

The review of internal spatial data bases is recommended to provide potential users with quality information prior to use. The review of provisional spatial data bases as described in this section is required and shall consist of two colleague reviews, one of which should be outside of the office originating the data set. A digital copy of the data base (or excerpt), evidence of these reviews, and the author's correction of any deficiencies shall be forwarded to the Branch of Scientific Publications (BSP) for identification as a publicly-accessible spatial data base. It is envisioned that these data bases would be identified in a published list of publicly accessible data bases.

Spatial data sets for which Director's approval is sought shall also provide a quality assurance plan to include the following information:

Identity of Data Base Administrator
Data revision plan including:

- the receipt and processing of updates
- frequency of data update
- update review procedures
- distribution and notification procedures

The quality assurance plan and the review materials will be reviewed by the BSP. Approved quality assurance plans will be released through the Open-File Series.

Publication of approved digital data sets

Data sets which have completed the review and approval process and are classified as either provisional or Director approved will be published in the Open-File Series. Such data must be made available to the public on a digital distribution medium appropriate to the size of the data set, the number of copies to be distributed, and the applications that will use it. Magnetic tape, floppy disk, and CD-ROM are examples of possible media to be used. Data may also be made available on-line, accessible through anonymous ftp or Wide-Area Information Server software.

Published digital data sets can be referenced using a citation similar to that for any other Open-File Report. The author is the person responsible for the creation or reprocessing and documentation of the data set. The title of the data set must include the source of the information used in the data set, a scale of reference, and a description of geographic coverage. The date used in the publication is the date associated with the preparation and release of the data and not necessarily the date of the source material. The media and data transfer format of the publication shall also be described in the citation. A data set version number shall also be provided by the author to track contents and updates.

Management and dissemination of approved digital spatial data

The Distributed Spatial Data Library (DSDL) is the collective set of local spatial data libraries in WRD that will house spatial data and data indexes for common use. Placement of standardized and reviewed data in DSDL will permit access to local data, remote data libraries and indexes, and will provide a mechanism for the automated retrieval of published spatial data sets across the wide area network. Documented digital spatial data sets must be published on a digital medium suitable for distribution or placed on-line to provide automated access. On-line indexes of approved digital spatial data sets will be developed to refer prospective users either to the on-line location of the data or to where a copy of the distribution medium can be obtained.

Policy for the accessibility and cost of access to on-line digital spatial data has yet to be defined for the U. S. Geological Survey. In the interim, the local office may elect to place published digital spatial data on-line for local and remote access to encourage its second-use potential.

Update of spatial data bases

The documentation for a spatial data set, as managed by the DOCUMENT program, includes a provision in the narrative section to keep track of revisions to a data set. Any revisions to published spatial data sets--or to data sets that are expected to become provisional or Director-approved--shall be recorded in the documentation of the data set, and submitted through a colleague review process. All revisions to provisional or Director-approved data sets will also be described in a separate revisions file which will list the data set name, the version number, the type of revision made, the identity of the party responsible for the revision, and the date of the revision. The revision information stored by the DOCUMENT program will be accessible on-line to all DSDL users.

SUMMARY AND CONCLUSIONS

The conceptual design of the Distributed Spatial Data Library (DSDL) is being implemented to serve the digital spatial data needs of the Water Resources Division through the use of public-domain data discovery programs and integration with existing GIS software. Spatial data, detailed documentation, data dictionary, and GIS-related programs are all being made available through the DSDL project to encourage re-use of data and applications programs within the Division.

Technologically, the solutions presented by the DSDL project are easily achievable with an understanding of the inner workings of a GIS software package and GIS techniques are used in the field. Without the necessary policy framework, also described in this report, the implementation of the DSDL project would be a greatly diminished success. Inasmuch as the DSDL model is being reviewed as a prototype by the Federal Geographic Data Committee and is being reviewed by the U.S. Environmental Protection Agency, the technical and policy framework are achieving more widespread recognition as a model spatial data management system.

REFERENCES CITED

Environmental Systems Research Institute, 1991, Using map libraries, ARC/INFO User's Guide version 6.0, July 1992 version.

National Institute of Standards and Technology, 1992, Spatial data transfer standard, Federal Information Processing Standard Publication 173.

U.S. Geological Survey, 1991. System requirements specification for the U.S. Geological Survey's National Water Information System II, S.B. Mathey, editor, Open-File Report 91-525. 622 pages.

*U.S. G.P.O.:1994-301-077:20