A Data Warehouse Digital Archiving Case Study at the North Carolina Department of Transportation

William Rasdorf, Ph.D., P.E.
North Carolina State University
Department of Civil Engineering
NCSU Campus Box 7908
Raleigh, NC 27695
phone: (919) 515-7637
fax: (919) 515-7908
email: rasdorf@eos.ncsu.edu

Kent Taylor, P.E.
North Carolina Department of Transportation
Traffic Survey Unit
Statewide Planning Branch
1554 Mail Service Center
Raleigh, NC 27699-1544
phone: (919) 733-9770 ext. 266
fax: (919) 715-7203
email: kltaylor@dot.state.nc.us

Larry Wikoff,
North Carolina Department of Transportation
Traffic Survey Unit
Statewide Planning Branch
1554 Mail Service Center
Raleigh, NC 27699-1544
phone: (919) 733-9770 ext. 268
fax: (919) 715-7203
email: wikoff@dot.state.nc.us

Key Words: data warehouse
digital archive
paperless office
archives
scanning
electronic documents
document management

Number of Words: 7413

A Data Warehouse Digital Archiving Case Study at the NCDOT

by
WILLIAM RASDORF, North Carolina State University,
KENT TAYLOR, NC Department of Transportation, and
LARRY WIKOFF, NC Department of Transportation

ABSTRACT

Data at Departments of Transportation are numerous and varied. Large legacy data stores exist in many forms, are stored in a variety of media, and include such diverse items as text (calculations, forms, correspondence, specifications, etc.), drawings, and photographs. Enormous collections of new data are continuously being created. Nearly all of the legacy data is in paper form. Even much of the newly collected data begins on paper, although most new data collection efforts are being converted to digital form.

Given the current state, the key problems faced by Departments of Transportation are how to find and acquire needed data, how to store it (warehouse), how to convert it to a useful form, how to relate it to other data (so that new analysis can be performed), how to manage it, how to access and use it, and how to archive it for both predicted and unexpected future needs.

In another context, State DOTs are asking what to convert to digital form, what to do with data that is not in digital form, how to manage and access digital data warehouses, and how to maintain digital data in the face of continuous and drastic changes in hardware, software, and storage media. This paper addresses these questions in the context of a study conducted by the Traffic Survey Unit (TSU) of the North Carolina Department of Transportation (NCDOT) which led to some insightful and unexpected observations.

INTRODUCTION

Traffic Survey Units are well aware of the difficulty of collecting and processing data. Various machines and data collection devices utilize different software and data formats, field forces receive varying training, paper forms are mixed with digital records, and paper to digital conversion efforts produce a mix of data formats. At the same time memory prices and processing time (CPU/cycle) are falling precipitously. A digital future is emerging that is leading to transition pains (to new methods and procedures for collecting data digital) and conversion pains (existing paper to digital forms). But even more ominous is the observation that the digital world will mandate the collection of ever increasing amounts of data, promote more (and better) analysis, etc. all resulting in continually increasing workloads.

It is recognized that the storage of electronic files is a significant consideration for any business, technical unit, or organization that produces a significant volume of data. State DOTs have seen that data generation in electronic forms is ever increasing; that the need to manage the proliferation of data, office, business, and enterprise files is essential; and, that planning is needed to adequately prepare to meet future data needs and to effectively archive existing data.

Yet DOTs face constraints. Outside pressures strenuously push for an all digital world and suggest that it be achieved yesterday. This is not realistic. What is needed is a critical analysis of the applicability, utility, and reasonableness of an all digital approach to data in the context of the DOT setting where technology, human resources, and budget limitations may dictate an alternative or hybrid approach.

The TSU of the Statewide Planning Branch of NCDOT initiated a study with these considerations in mind. Its objectives were to:

1. Develop a set of guidelines to use to estimate server disk space requirements for several years, based on:
   - growth of data which is currently collected or processed, and
   - data that is not currently collected or processed in digital format but may be in the future.
2. Identify data that is not currently in digital form but which would benefit from being collected, processed, and stored digitally.
3. Establish guidelines for estimating storage requirements for digitally archiving data which is currently on paper and stored in file cabinets, such as reports and letters.

These original objectives were intended to support informed resource planning which would be adequate for use for several years, thus saving money by eliminating the need to make unnecessary hardware and software purchases and expenditures in the interim period.

KEY FINDINGS

The digital storage study was initiated in June of 1998 (1, 2). This study determined that the cost of converting large amounts of the Unit's existing paper (text, forms, and drawings) to a digital format would be substantial and probably not cost effective in most cases.

On the other hand, the storage of data and information that is already in a digital format is much easier and more efficient and a digital form of storage would, indeed, be employed for such data. The cost is relatively low, and the benefits of having such information archived outweigh the financial requirements. Still, the issue of archiving comes into play. Getting digital data is very different from subsequently storing and using it.

An additional key finding was that for the NCDOT TSU, 80 percent of all information management requests go back 3 years or less and only 20 percent of the requests fall between the range of 3 to 15 years (2). This suggests strongly that it is not cost effective or necessary to

archive information beyond 3 years in such a way that it is readily available to customers. Still, AASHTO recommends retaining data for 10 years (3).

With this in mind it is important that DOT's implement a formal set of procedures and practices for data collection, storage, and use to ensure constant, up-to-date, digital information and to avoid "back-logging" in the future. However, some organizations wait for information to accumulate and then must repeat the labor intensive, paper-to-digital conversion process. This neglectful approach will provide information that easily meets the needs of the 20 percent of users who read information beyond the previous 3 years, however, it does not provide the more current information that is needed 80 percent of the time.

Finally, when thinking about archiving and digital formats in general, significant consideration must be given to the archiving of application software (in addition to archiving data) and even the hardware platforms on which it operates (4). Too often software may be archived and saved, but it cannot be run at the future time that it is needed because the hardware that created it is no longer available. For example, information currently stored on 5.25" floppy disks will eventually be irretrievable because the drives that support these disks have become obsolete and are increasingly difficult to locate.

TRAFFIC SURVEY UNIT BACKGROUND

The primary function of the TSU is to collect transportation data and to format it in a meaningful manner. The unit also processes data to derive traffic and travel statistics that are of use to others. Approximately 90 percent of the data is specific to traffic volumes, vehicle types, and vehicle weights. The Unit may be viewed as a transportation data warehouse, which is one aspect of the overall function of producing traffic information. In the future, this Unit would like to provide on-line access to its data for all those who might be interested in traffic information.

Major Applications

The primary data collection activities of the TSU and a description of the general nature of the data are provided herein. The databases used to store survey data are described and the products generated by the Unit are as well.

Survey Programs

The TSU is responsible for maintaining data collected in a series of programs. In this context, ***program*** is defined as a data gathering and analysis activity. A general description of the TSU programs, followed by more specific details about each program, is given below.

> ***Continuous Count Program (ATR)*** – In this program, all vehicles are individually counted using machines that can be remotely accessed without personnel having to field check them. These machines are permanent counting stations. The data they collect is used to generate seasonal factors to determine Annual Average Daily Traffic (AADT) and to generate Design Hour Volume (DHV) information.

Coverage Count Program (PTC) – This program uses conventional vehicle counters that do not distinguish between vehicle types to record the amount of vehicle traffic at a particular location. Personnel utilize the counters for a short-duration count. The station locations are fixed, but the counters are portable and are installed at regular time intervals. This data is also used to determine Annual Average Daily Traffic (AADT).

Vehicle Classification Program – In this program, vehicles are counted and classified into 13 vehicle types for the purposes of traffic analysis and roadway design and improvement.

Weigh-in-Motion Program (WIM) – In this program, all types of vehicles are weighed using special scales placed in the roadway. The weight recorded is weight per axle.

Special Count Program – Often a department may require a specific type of count at a particular location as part of a special study or to design an improvement. In that case, a special count, beyond the scope of the other counts mentioned in this section, may be taken in order to provide the data for the special need.

Vehicle Occupancy – In this program, the number of occupants per vehicle is counted.

Origin and Destination Survey **(O & D)** – This program surveys drivers and passengers to determine where their trips originated from and what their terminal destination is. O&D surveys are also conducted to determine trip purposes.

Drawbridge Reports – In this program, the number and type of vehicles using a drawbridge is recorded. The number of drawbridge openings is also recorded. All files are maintained in paper format.

Ferry Utilization Reports - In this program, the number and type of vehicles using a ferry is recorded.

Program Details

- ***Continuous Count Program (ATR)*** – This is a program where traffic counts are continuously collected using Automated Traffic Recorders (ATR). Counts are provided hourly by lane continuously. There are approximately 100 of these recording stations throughout the state operating 365 days a year. Counts are normally recorded as number of vehicles per hour.

- ***Coverage Count Program (PTC)*** – This is a program where traffic counts are periodically collected using portable traffic counters. These Portable Traffic

Count (PTC) stations are generally placed at fixed locations. Approximately 58,000 of these stations are operated throughout NC. The stations are divided into several categories, including Urban, Primary, Paved and Soil Secondary, Interstate, and Interstate Ramps. All stations are counted annually (one count per year) with the exception of the Urban and Paved and Soil Secondary, which are counted biannually. Counts are normally taken in 48 to 72 hour intervals and are recorded as daily totals.

- ***Vehicle Classification Program*** – There are approximately 300 sites devoted to counting and classifying 13 vehicle types. Of these 300 sites, one hundred stations are utilized for counting each year. (Each count stations is operational once every three years). In addition, there are 350 truck-monitoring stations of which 300 are utilized to count annually. That is, 50 sites are rotated out of service each year and return to service the following year. These classification sites employ either manual or automatic classification devices. Manual classification data is collected using Jamar count boards. Machine classification data is collected with Peek ADR 1000 equipment by lane by hour.

- ***Weigh-in-Motion Program (WIM)*** – There are 23 SHRP sites and 5 WIM Monitoring Stations that are in continual operation.. WIM data is also used to compute ESAL (equivalent single axle loads), which is used by the Pavement Management Unit. WIM also monitors the number of vehicles that are in violation of the Federal Bridge Formula.

- ***Special Count Program*** – A special request usually occurs when there is a specific need that cannot be met using the data collected in other Traffic Survey Programs. By a special request from another Unit within NCDOT, intersection studies (turning movements), daily and hourly volume counts, vehicle classification counts, travel time studies, and other vehicle data collection activities may be conducted.

- ***Origin and Destination Survey*** – This is data related to the origin and destination of travelers. Data is collected manually by individuals asking specific questions of drivers and passengers. This type of survey is conducted periodically at the discretion of the department.

- ***Ferry Utilization Reports*** – Data for ferries contains counts and classifications of vehicles using the ferry system and accounting for vehicles towing trailers.

The TSU also stores old maps (going back over 50 years) that include information such as Annual Average Daily Traffic (AADT) and Design Hour Volumes (DHV). These are stored in various filing cabinets.

Databases

It should be pointed out that the TSU maintains a tremendous amount of data. A series of data inputs and outputs are generated from the programs described above. In most cases, the data inputs are considered “raw” data (i.e. the data obtained from the counting stations). This raw data is then "screened" (commonly referred to as "edited" or "adjusted") to screen out bad data and to correct errors in the data. The screened data must then be analyzed and processed into meaningful information.

For the most part, only processed data is included in the TSU databases. This processed data is much more useful to other units within NCDOT. Raw data cannot be used effectively until it is processed to yield information. The following table summarizes raw data storage collection methods, storage locations, and storage types used within the Unit.

Paper Field Sheets	**Collection Method**	**Storage Location**	**Storage Type**
Continuous Count	Digital	Mainframe	Text
	Digital	Server	Binary & Text
	Digital	File Cabinet	Paper
Coverage Count	Field Sheet	File Cabinet	Paper
	Field Sheet	Server	Access DB
	Field Sheet	Mainframe	Text
Vehicle Classification	Digital	Server	Paradox DB
	Digital	File Cabinet	Paper
	Digital	Diskette	Text
Weight in Motion	Digital	Local/CD	Binary, Text, & VTRIS DB
Turning Movements	Field Sheet	File Cabinet	Paper
	Digital	Server	Lotus SS
Travel Times	Digital	Server	Text
Vehicle Occupancy	Field Sheet	File Cabinet	Paper
	Field Sheets	Server	Access DB
Origin Destination	Field Sheets	File Cabinet	Paper
	Field Sheets	Server	Text/SS
Drawbridge	Field Sheets	File Cabinet	Paper
	Field Sheets	Server	Quattro SS
Ferry	Field Sheets	File Cabinet	Paper
	Field Sheets	Server	Quattro SS

The field sheets are the actual paper sheets filled out in the field by the technician. Note the diversity in storage locations including an IBM mainframe, a local file server, diskettes, and CD's for digital records storage and filing cabinets for all of the paper. The methods of data storage include paper, spreadsheets (Quattro, Lotus, and Excel), flat files, and databases (Paradox, Access).

Products

Processed data is distributed through two key products - maps and reports. Reports are textual documents containing voluminous tables of traffic count data and include monthly ATR reports, quarterly WIM reports, and annual ATR, PTC, classification, and WIM reports.

Two sets of base maps are currently produced. One set identifies the actual count stations and their location. The second set records the processed annual average daily traffic volumes. It provides a graphical representation of traffic volumes on roads throughout the state. The majority of these maps originate from the GIS Unit's Mapping Section where they have been digitized into, and printed from, MicroStation, a CAD package (6). The traffic counts are handwritten on the hard copy maps. The maps are then reproduced and distributed to users.

Soon, these maps will be accessible in an on-line format, where traffic volumes can be more quickly ascertained. The maps will originate from ArcInfo, a GIS package (7). The traffic counts will be entered as coverages in ArcInfo and subsequently made available on the internet. To enable this to happen the Unit is redesigning and rebuilding its databases and linking them with the new NCDOT base linear referencing system. Initial tests at NCDOT have demonstrated that large volumes of data can then be viewed either in tabular or graphical form and that spatial data queries can be made either from the DBMS or from the GIS (8).

Data Issues and Obstacles

State DOTs face numerous serious problems with respect to data and information management (9). Many of these have a direct impact on the structure and nature of a DOTs data warehouse - the collection of all data collected, stored, maintained, and used by the department in its day-to-day operations. This, in turn, has an impact on how and where that data is archived.

The TSU has determined that its current data management system is cumbersome and difficult to update and use (8). As noted above, and as is typical for all DOTs, there are many different data systems that are contained in many different database types. These different data systems and database types must be consolidated. Data downloaded from the mainframe and data acquired from other Unit's databases has led to replication of data and parallel databases. This can result in outdated data because the original databases are out of the owner's control. Therefore, data can no longer be updated and maintained properly. Furthermore, there is no uniform mechanism for making changes to these parallel copies of the databases. Obviously, this is not the most efficient means of data management. The database types and systems must be consolidated to simplify access and retrieval.

Clearly it would be advantageous to restructure the entire data management process to allow easier access to data. In order to do this, software must be standardized by selecting database management systems that can be integrated with one another. Data queries that are both flexible and comprehensive must be possible so that specific information can be obtained when needed and in a proper format. The Unit would like to have a standard structure for all reports and queries.

Reports need to be designed with a higher quality and efficiency (i.e. improve readability and eliminate redundant information). One way to increase report efficiency is by delivering reports to customers and users in an electronic format at periodic intervals (i.e. monthly, yearly, or on-demand). This means information should be available on-line without ever having to see a printed hard copy.

Presently, the TSU is posting scanned traffic volume map images on the internet for agency and public use. In this manner data is being made available to users and customers while reducing the time taken by DOT personnel to supply it. The unit is now moving toward utilization of internet-based GIS software to further reduce the amount of direct interaction with the general public while providing public access to the data in both a tabular form and an easy to review spatial form as well. In this manner users will not only be able to see and access data, but they will be able to query it and perform limited analysis as well.

To achieve these goals a design for a new database for the Coverage Count Program has been completed (10) and is being implemented using Oracle. Significant data sharing with inside data users and outside data users is expected (11, 12). The system will have a seamless interface between Oracle and ArcView, allowing graphic display of all traffic data.

All of the issues and obstacles mentioned here point to one basic need - to have data in digital form. Before discussions can proceed about which machine to store data on, before discussions can proceed about which software to process data with, before plans can be implemented to collect more data, and before the format of the data can be discussed, the data must be in digital form. This results in large data warehouses. The following section discusses issues related to archiving large volumes of data and information and also addresses the conversion of paper data and information resources to digital form. Understanding the issues related to conversion and archiving is critical to developing a strategic plan for dealing with enterprise data, be it legacy, current, or future data and information.

PAPER STORAGE

Currently, the TSU has thousands of paper documents stored in filing cabinets. The scanning of these documents (and their storage as digital files) would eliminate the need for the paper files. Therefore, if the paper is scanned into a database and can be easily accessed, the original paper files could then be destroyed or recycled. However, there may be some documents and maps that cannot, for various reasons, be scanned into digitally readable files. These documents should then continue to be filed in cabinets. And, if no scanning takes place, existing hard copy document archives must be maintained regardless.

But, no matter whether hard copy or digital archives are maintained, an electronic form of indexing should be used to locate documents regardless of the form in which they are stored. Yet electronic indexing for paper files, to support finding documents (and the information they contain), can sometimes pose an interesting problem. Such documents may be stored in widely distributed locations. Even if stored in a central location a mechanism must be provided for

accessing those documents based on some useful and easily recognizable indexing scheme. This indexing scheme should be implemented using a database that can easily and readily be searched using various index categories and levels. The user is then provided with a precise description of the physical storage location of the document they are searching for.

Thus, organizational units face hard choices about how much paper to convert to digital form. Why are these choices so hard? Because they require capital outlay, they require knowledgeable personnel, they require management support and commitment, and they depend on rapidly changing and unstable technology. As a result there is concern about exactly what to do. This paper makes some helpful suggestions that might significantly benefit State DOTs in this regard.

ARCHIVING

Archiving addresses a key question in any organizational unit's transformation from a paper-based records system to a digital-based records system - "What must be stored?" Project based data includes text (correspondence, meeting minutes, fax, transmittals, etc.), drawings, photographs, and more. Business operations including marketing, personnel, payroll, and others generate information that must be maintained as well. Finally, software itself, of all types (engineering, technical, and operations) must also be considered in the context of storage and archiving.

There are other important questions. How much of the ***past,*** existing paper should be converted to digital form and included in the new digital store? This question assumes that the decision has already been made to convert all ***current*** and ***future*** data to digital form! Furthermore, it assumes that the organization is also already doing so! The point here is that the present and the future must be resolved and functioning well before it makes sense to go back, work with, and make conversions of, previous data, information, records, and documents.

Scanning Existing Documents/Drawings

Currently, the only feasible way to convert a paper document/drawing into an electronic form is through scanning. One issue which quickly arises is who will do the scanning. There are two primary options to consider for document scanning. The first is to hire an outside source, referred to as a Service Bureau. The second is to scan the documents using internal personnel. Depending on an organization's scanning needs, one of the two options will produce a more feasible outcome.

A second issue that arises in converting paper documents into retrievable digital forms and drawings is what should be scanned. The following discussion includes three key topics concerning scanning (1) general scanning information that pertains to both 8.5" x 11" documents and standard drawings (maps), (2) detailed information dealing only with 8.5" x 11" documents, and (3) detailed information dealing only with standard engineering drawings and maps.

Outside Source Considerations

Outside sources are often referred to as Service Bureaus. The cost for hiring a Service Bureau varies considerably depending on whether a document/drawing is bound or unbound, well preserved or poorly preserved, and whether there are many pages compared to very few.

Whether the bulk of scanning is bound or unbound will have a substantial affect on the cost of scanning. A bound document requires someone to manually scan each page using the flatbed scanner rather than the automatic feeder. This manual feeding increases the amount of man-hours required which, in turn, increases the cost/page. Alternatively, the document may need to be disassembled, scanned, and then rebound.

Another important factor is the condition of a document. A document in poor condition often cannot be fed through the automatic document feed portion of the scanner due to the risk of damaging the document. In such cases, the documents must be individually laid on the flatbed portion of the scanner to perform the scan. This manual feeding increases the amount of man-hours required which, in turn, increases the cost/page. In addition, the legibility of the scanned image of a document in poor condition may be questionable.

One of the most important factors to consider is the number of documents to be scanned and the total quantity of pages. Most Service Bureaus prefer to deal in bulk. The more pages scanned, the less it costs per page.

On average the approximate cost range for 8.5” x 11” documents may be anywhere from $.05 to $.80. The low range ($.05) is generally the price if millions of documents need to be scanned and the high range ($.80) is for copying relatively few documents that are in poor condition or are bound or both. Thus, cost of hiring a Service Bureau varies considerably and can be an expensive solution.

Inside Source Considerations

Sometimes the best solution is to perform the scanning internally. In order to choose which solution is more cost effective, one must consider the hardware and software needed for the task. For an inside source solution, all of hardware, software, and labor factor into the overall cost in addition to all of the production considerations mentioned in the last section regarding document size, condition, and volume. But the key to finding the most efficient and cost effective hardware is estimating how many documents must be scanned. From this data the equipment required to get the job done can be determined.

Small to moderate 8 1/2" x 11" scanning jobs require scanners ranging in price from a low of $1000 and up (Fujitsu Scanpartner 10C - 10 ppm ADF Scanner - approximately $1200.00, Fujitsu 3093GX - 27ppm ADF Scanner - approximately $3500.00). These scanners are designed to be used in a document archiving environment, and are capable of scanning hundreds of documents per hour for extended periods of time.

A variety of software exists for document archiving and management. Axxis, for example, is imaging software that creates a "picture" of the scanned document and stores it in CCITT-4 standard fax format (13). During this process, the software utilizes a modified OCR (optical character recognition) technology to extract words within the document and create a database from these extracted words. Additionally, this software allows one to place a hierarchical classification format on all documents scanned. The database may then be searched by using the hierarchical classification system, or by entering any word or phrase one is searching for.

The price for an Axxis single user system is $5,500. If the software needs to be accessed over a network, the price varies depending on the number of concurrent users needing to access the stored files. The base price for the network software is $7,000 for the server software, and requires at least two concurrent user licenses to be purchased at a price of $2,000 each. The price per concurrent user remains the same for up to 5 users and then begins to drop for every user thereafter. In summary, a five concurrent user package would cost approximately $17,000. Thus, the new hardware and software that would be required to scan and store 8 1/2" x 11" documents would require a capital outlay of $20,000 and does not include the cost of actually doing the scanning.

Image Storage Requirements

A key concern in storing electronic data is the amount of space that image files require. In the overall scheme of things the document management software requires very little in storage space; the controlling factor is the number of documents scanned and stored on the chosen media. A typical 650 MB CD can hold 40,000 to 60,000 images. A standard filing cabinet can hold an average of 14,000 pages. Therefore, about four filing cabinets can be consolidated onto one CD.

Total Document Image Cost Estimate

The NCDOT TSU developed a detailed cost estimate for 8 1/2" x 11" documents. In pure dollars its findings indicate that there is a distinct advantage to doing the paper-to-digital conversion work internally for small numbers of documents (< 50,000). First, there is an overall savings when compared to utilizing a service bureau, and second, a scanner will be in place to scan future documents. These findings include the cost of purchasing the hardware and software and the labor involved in doing the scanning. One of the reasons the cost is higher for Service Bureau work is that even though someone else is doing the physical scanning, the software must still be purchased to access and view the documents after they are scanned. Thus, the investment cost for new hardware is the primary difference between the two choices and it is small, comparatively.

Scanning Standard Drawings/Maps

The basic cost of scanning standard 24" x 36" maps and drawings involves a much higher cost than scanning an 8.5" x 11" letter-sized document. This cost may range from $1.50 to $3.00 per sheet, on average. In addition to the previously stated conditions of the documents, this variation in cost is due to several additional factors such as size, line art versus grayscale, file storage type, and scan resolution.

Although purchasing hardware for inside source map scanning is a consideration, the cost of this purchase will most likely be more than the cost of hiring a service bureau, unless a sizeable number of documents are to be scanned. Additionally, a service bureau can provide the expertise needed to provide clear images while minimizing the storage space required. Unequivocally, for the document scanning load considered by the NC TSU it became clear that, for larger documents and maps, purchasing scanning equipment deserved little to no consideration as it was clearly not cost justified.

Technical Support Costs

Software will need to be acquired to organize, input, store, manage, access, and monitor the increasing collection of data. Most often, some form of software (with limited capability) is included in the off-the-shelf purchase price of such items as scanners, digital cameras, CD writers, and the like. However, if the software does need to be purchased, the price may have a significant impact on the total cost of storing large amounts of information. Still, hardware and media costs will be an important factor when deciding to archive project information.

Technical support is clearly necessary in a computing world that is becoming increasingly sophisticated and specialized. A designated person or department must be identified who will assume responsibility for managing the information system of the entire Unit. Determining the organization, scope, and nature of the technical support function is the dilemma faced by DOTs. What has become clear at NCDOT is that hardware and software technical support personnel must be housed directly in most units, in addition to having a centrally responsible hardware/software acquisition/training/development group. This individual(s) needs to be knowledgeable in computing as well as in the discipline-oriented activity of the specific unit. They can then provide useful, focused, and timely support while drawing upon the expertise of the central organization if and when needed.

FILE MANAGEMENT

File Management acknowledges that the proliferation of digital data, including thousands of data collection files, different versions of files, different versions of software, etc., creates the potential for a chaotic bin full of files. Such a collection of files may have lost their meaning to a user, are inaccessible, or are not recognizable over time. The need to develop a naming convention for files, to establish an organizational structure for the storage of files, to set standards by which all persons use and store files, and to appoint a file administrator is obvious.

Currently, the TSU has a plethora of software and data. Both data and program files need to be easily accessible and retrievable. The key to retrieving them in a fully digital environment is proper management of the files.

Access to Information

The acquisition of data will only increase (most likely at an exponential rate) over time. Access to archived project information is vital to any DOT's competitiveness, success, and ability to fulfill its mission. As the number of employees and projects increase, many project managers and project engineers find it increasingly necessary to use information from previous projects as a database for current ones. Additionally, new employees may find it useful to browse recent projects to become aware of the methods and procedures and the documents they use. Information is becoming an increasingly vital resource.

When speaking of access to information the following need to be considered: time of access, reliability, and management. A brief discussion of access time and reliability is appropriate here. While there exist many types of data storage media, they vary greatly in terms of their ability to readily access information, their reliability, and their life span. In the NCDOT TSU most projects, upon final completion, are archived for future use so as not to overburden on-line network storage. The current method utilizes CD-ROM technology. There were three types of electronic storage that were initially considered by the TSU (2).

On-Line network storage gives all workstation users direct access to network hard drives which contain all project information for current projects. Access is instantaneous.

Near-Line storage utilizes optical discs that store any file that has not been modified for a certain length of time (a time stamp is sometimes utilized for this task). Each workstation has access to a CD "juke box" where archived files can be accessed relatively quickly (within several seconds, as opposed to immediately) and copied again to on-line network drives.

Off-Line storage uses CD-ROMs to store all project information for completed projects. Access is more difficult and timely, but a well organized filing system will allow an information systems manager to easily locate any file within several minutes to an hour.

These storage methods address archived data storage. Management is discussed below and encompasses both the management of archived data as well as the management of currently active data sets that reside on PC hard drives and on the server.

Managing Information

It is apparent that data management requires an organizational system that will not only ease data access, but limit it to appropriate personnel. Computer files are usually stored in a hierarchy of folders of directories, which have the benefit of imitating what many DOTs currently do with paper files. Therefore, locating certain electronic files becomes a simple translation from the paper world to the digital world. As data is archived, a directory tree structure can be created. Optimal storage requires that employees purge project directories before archiving in order to eliminate duplicate data and establish a more efficient filing system.

Managing this growing file structure will become more important as time goes by and will require a focused and dedicated effort by the system manager or responsible individual (4). For various reasons, certain people will need to be denied access to data owned by the organization or unit and other data will need to be readily available for all to use. A password protection system must be implemented, but this will need to be controlled by the system manager and a system of checks and balances should be put in place to ensure adequate control. If information is to be made available to the outside via the world wide web on a project by project basis, or for use by different units, further management will be imperative.

Application Software Archives

One critical item that must also be considered is the archiving of applications software. A careful organization wide plan should be developed for archiving its most commonly used software. This is especially true for the software that manages the archived data. If that software is rendered useless in the future, the data becomes inaccessible and its value and utility are crippled.

On a consistent basis software is upgraded and changed by its originator and new versions are issued. It is critical that old versions not be discarded under the assumption that the new versions can always run using older data files. Rather, older data files must be carefully tagged so that they indicate the software version for which they are suited and that version of the entire software application should be archived along with the data, in compliance with licensing agreements.

CONCLUSIONS AND RECOMMENDATIONS

This paper raises questions. It is clear that there are two topics of issue and concern to the NCDOT TSU. One is the question of how much disk space should be acquired to hold all of the Unit's digital information. The second question is related to the transformation of the Unit to a "paperless office" mode of operation and deals with how much information should even be converted to digital form. This paper seeks to comment on that transformation.

One of the more useful observations was that the known space requirements for programs and data (as determined by a detailed study), both locally and on the server, was not unexpectedly large. That is, current digital requirements can easily be met through current disk technology. Still, what is in order is a cleaning up of older version of files, a garbage collection effort if you will, to free up space currently being used by duplicate files.

Regardless of the above noted discrepancy, we were able to determine both the ***needed*** space for programs and data and the ***used*** space. What the numbers tell us is that the overall space requirements are not projected to be too high. The amount currently used is on the order of 12GB. Given that this amount of information can easily be stored on the hard drive of most newly purchased computers there really isn't a storage problem of a serious nature. For a mere $2500 one could purchase a new computer, move all software and data to it, and use that computer as the archive. The advantage of doing so is that all the software would be operable and readily available for use if needed, as would the data. Furthermore, if it is connected to a network, all data could easily be shared as well.

Another key conclusion is that a formal, documented file management system for the data is critical. This would include the establishment of formal standards for operational procedures. This would obviously help with respect to disk storage space but it has the potential to do more. It would help to establish a standardized way for the Unit to define, work with, format, and archive its data. In turn, this would increase accuracy, promote information exchange, and enhance the access to the data by outside users. It would standardize the names of files, provide a means of storing meta data about the files themselves (who is responsible for the file, where it came from, who is using it currently, who has duplicate data, etc.), and assist the Unit in dealing with archived data.

What the above two conclusions (and in a sense, recommendations) articulate is a point that many organizations are beginning to recognize - that transportation organizational units need to consider formally including the concept of "information manager" in their structure. That is, they must either incorporate a new organizational unit job slot, charge an existing employee with new responsibilities and rewrite their job description, or have this service provided by an information technology support unit. When the new job slot is not an option, the employee designation option needs to be considered especially carefully. For a unit who's primarily responsibility is to deal with information and data, this is a serious consideration. The closer the computing expertise resides to the unit the more successful its automation and information technology efforts will be.

It is clear that the focus of engineering data storage thinking should be on the future rather on the past – at least for now. That is, efforts should be made to plan an evolutionary path to generating all new data and information digitally. This should be done so in a carefully considered and step by step manner to ensure that the transition to a digital office is successful. Thus, for the NCDOT TSU, it was determined that existing archived paper files be maintained for the foreseeable future and that no digital archive would be created for them. Instead, it was recommended that all automation efforts concentrate on getting the unit into a position of collecting all new data digitally. This means that, as time goes on, the amount of new paper that is generated will be reduced, and eventually very little new paper will be produced. It was deemed unwise to focus on converting old paper to digital files while continuously generating new paper. Rather, if new paper generation can be reduced or eliminated, then consideration can be given to the existing paper conversion issue.

SUMMARY

Traffic Survey Units in particular are data and information intensive organizational entities. They must perform field data collection using vendor supplied equipment and proprietary software (or even hand written forms). They store, access, and process the data using paper, databases, files, and maps. In distant years past this was all done on paper. Today this is mostly done on paper in combination with digital files.

As the evolution to ever increasing levels of automation continues it is important to examine the quantity of data one has and the size of the warehouse needed to store it. This paper suggests that quantity does not pose as much of a problem as one might think, and that storage will not be

a problem in terms of capacity or cost. Still, concerns about rapidly changing media must be examined to ensure data access over long periods of time.

It is also important to consider to what extent it makes sense to go digital. Our findings indicate that for new data one should evolve to a completely digital system over a timeframe that is reasonable and realistic for an individual DOT. The caveat is that this requires standardization and the development of formal procedures for naming, referencing, and location. One should use GIS to store spatial data. One should use a relational database to store attribute data. Finally, data should be made available over the internet to obtain broader distribution and empowerment of customers and other interested parties in its use.

REFERENCES

1. Rasdorf, W. J. *Projecting Future Electronic Storage Requirements Within the Client/Server*. Technical Report, NCDOT, June 1998.

2. Rasdorf, W. J. *A Digital Information Storage Study at NCDOT*. Technical Report, NCDOT, August 1998.

3. Joint Task Force on Traffic Monitoring Standards of the AASHTO Highway Subcommittee on Traffic Engineering. American Association of State Highway and Transportation Officials, AASHTO Guidelines for Traffic Data Programs, Washington, DC 1992.

4. Mooney, R. W. A Call for Quality in Engineering Business (And the Computer Tools to Get Us There). *Journal of Computing in Civil Engineering*, American Society of Civil Engineers, Volume 9, Number 3, June 1995, Pages 191-193.

5. Kiel, D., Rasdorf W., Shuller, E. and R. Poole. "An LRS Model for NCDOT," *Transportation Research Record*, Number 1660, National Research Council, Washington, D.C., 1999, Pages 108-113.

6 Bentley Engineering. *MicroStation*. Exton, Pennsylvania, 1999.

7. Environmental Systems Research Institute, Inc. *Understanding GIS: The ARC/INFO Method*. Redlands, California, 1994.

8. Rasdorf, W. J. *The Conceptual Universe Database Schema: Design Issues and Decisions*. Technical Report, NCDOT, August 1999.

9. Rasdorf, W., Shuller, E., Poole, R., Abudayyeh, O., and F. Robson. Information Management at State Highway Departments: Issues and Needs. *Journal of Transportation Engineering*, American Society of Civil Engineers, Volume 126, Number 2, March/April 2000, Pages 134-142.

10. Rasdorf, W. J. *A Spatial and Attribute Database Schema Design for Traffic Survey Data*.

Technical Report, NCDOT, December 1999.

11. Bower, S. T. and B. Harris. Linear Referencing in a Data Warehousing Environment. In *Proceedings of the 1998 Geographic Information Systems for Transportation (GIS-T) Symposium*, Salt Lake City, Utah, April 1998, Pages 86-95.

12. Dubois, R., Armentrout, N., and P. O'Packi. TIDE: A GIS-Linked Data Warehousing Approach for Building an Integrated Transportation Information Environment. *Proceedings of the 1998 Geographic Information Systems for Transportation (GIS-T) Symposium*, Salt Lake City, Utah, April 1998, Pages 68-85.

13. Automated Office Systems of NC. *Axxis Scanning Software Manual*. Raleigh, North Carolina, 1997.