Digital Engineering Document Repository Optimization

By:
David P. Hale, Ph.D.,
Area of Management Information Systems
The University of Alabama
Tuscaloosa, Alabama

and

Andrew J. Graettinger, Ph.D. and Laith Alfaqih
Department of Civil, Construction,
and Environmental Engineering
The University of Alabama
Tuscaloosa, Alabama

Prepared by

AISCE

Aging Infrastructure Systems Center of Excellence
The University of Alabama

And

UTCA

University Transportation Center for Alabama
The University of Alabama, The University of Alabama at Birmingham,
and The University of Alabama in Huntsville

AISCE Report Number 071001
UTCA Report Number 05106
October 4, 2007

Technical Report Documentation Page

1. Report No 05106	2. Government Accession No.	3. Recipient Catalog No.

4. Title and Subtitle Digital Engineering Document Repository Optimization	5. Report Date October 4, 2007
	6. Performing Organization Code
7. Authors Dr. David Hale, Dr. Andrew Graettinger, Laith Al-Faqih, and Carlos Sanchez	8. Performing Organization Report No. UTCA Report 05106 AISCE Report 07091
9. Performing Organization Name and Address AISCE -106 Bevill 201 7th Avenue The University of Alabama Tuscaloosa, AL 35487-0208 205 348 5525; aisce@ua.edu	10. Work Unit No.
	11. Contract or Grant No. UTCA 05106 AISCE Report 071001
12. Sponsoring Agency Name and Address University Transportation Center for Alabama Department of Civil and Environmental Engineering University of Alabama Box 870205, Tuscaloosa, AL 35487-0205	13. Type of Report and Period Covered Final Report: December 15, 2005 – June 28, 2007
	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Information technology now makes it possible to create digital repositories of previous hard-copy document archives. The benefit of such a system is to "have data at our fingertips no matter where we are and what document contains the data". However such a benefit presumes easy document retrieval. Prototypes of such systems indicate that as point solutions, rapid direct retrieval of documents can be achieved. But as the diversity of documents, user groups, and desired retrieval outcomes increases, the ability to retrieve "the" document and its associated data becomes less likely. Instead, lists of documents are frequently found and navigation through these lists is often clumsy at best. Safety concerns require that existing documents be readily available when needed, that such documents are not lost, and that version and ownership control exists.

The University of Alabama's Civil Engineering and Management Information Systems Departments, and The Alabama Department of Transportation (ALDOT) will collaborate on this project to research the issues and prescribe effective approaches for synthesizing leading practices from existing transportation organizations and the information systems/sciences discipline.

17. Key Words Document Repository, Archiving, Transportation, Document Management		18. Distribution Statement	
19. Security Class (of report) Unclassified	20. Security Class. (Of page)	21. No of Pages 85	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Table of Contents

List of Figures

Executive Summary

This project provides guidance to transportation organizations as they create digital repositories of their technical engineering documents. The goal of the project was to refine digital storage structures beyond that of hardcopy folders, and to refine retrieval capabilities beyond smart PDF capabilities. To achieve this goal, the project:

1. Conducted a literature review of transportation specific document management systems;
2. Conducted a review of current Information System/Computer Science (IS/CS) literature concerning advanced and evolving abstraction techniques and meta tag structures;
3. Benchmarked state department of transportation for leading practices and unresolved issues;
4. Developed advanced requirements for ongoing and/or historic document acquisition strategies;
5. Developed advanced requirements for document retrieval strategies; and
6. Developed alternative conceptual design for storing and retrieving documents.

Section 1.0 Document Management in Civil and Transportation Engineering

1.1. Introduction

Transportation systems are complex engineered facilities with vast amounts of supporting data and documentation. The amount of information produced and processed on a daily basis at a Department of Transportation (DOT) is becoming beyond the capability and capacity of current ad-hoc document management systems. Accurate, reliable, and readily available data is an essential tool for managing and improving transportation systems. Transportation professionals must find efficient ways to meet federal, state, and municipal needs for modeling, data management, traffic counts, and performance measuring (Tate-Glass et al., 1999).

In a globally competitive environment, electronic data management (EDM) technologies give businesses an advantage. EDM technologies ease the process of collecting, accessing, analyzing, protecting, and managing data. EDM results in an integrated organization by impacting work flow, required resources, productivity, project duration, and cost (Back et al., 1995).

Intelligent Transportation Systems (ITS) have emerged that provide data related to: traffic surveillance and control, incident and emergency management, public transit, vehicle crash, commercial vehicle operation, environment and weather conditions. These automated data collection tools are preferred because they reduce bias and support the creation of system performance measures (Liu et al., 2002); however, ITS generated data is massive and complex, requiring an archiving system capable of managing the data efficiently. Any such system should be compact, able to quickly and securely retrieve large amounts of data, and be compatible with various computer platforms. Ideally, the system would be supported by analysis tools, be capable of self-describing data, and have low operating costs (Kwon et al., 2003).

In addition to real time data, DOTs are struggling with historic documents addressing issues such as engineering standards, facility construction, and facility maintenance. Currently these documents are in a paper format and organized with an ad-hoc cataloging system that is prone to loss and errors. Sustainable, quick and accurate access to plan sheets and reports is essential to support transportation engineering decision makers. This chapter reviews published literature related to digital document management issues faced by civil engineers and more specifically transportation engineers. Currently used software as well as case studies is reviewed. Successful implementation of document management systems in the construction industry is also presented.

1.2. Data warehousing

The storage as well as retrieval of digital data is known as data warehousing. Data warehousing is a technique that provides an information architecture that can serve as an enterprise-wide source of data for performance analysis and organizational reports (Liu, 2002). A data warehouse is a global read only analytical database that is used as the foundation of decision support systems (Poe et al., 1998). Data warehouses are designed to be accessed quickly through queries, thereby enabling effective and rapid decision making (Chau et al., 2002) by delivering understandable information in a usable business context (Liu et al., 2002). Data warehousing was established more than a decade ago by private industry to increase market share and productivity (Papiernik et al., 2000). A data warehouse is not simply a place to dump data. It is not a single technology and may include electronic data or microfilm. It is also not a panacea for all agency data problems (O'Packi and Lewis, 1998). Data warehousing involves organizations extracting value from their data assets (O'Packi and Lewis, 1998). The heart of a potential transportation data warehouse lies in the warehouse itself. The warehouse should be comprised of the following components (Papiernik et al., 2000):

- integrated core of transportation data organized according to its subject content,
- a suite of data service processes that acquire, integrate, prepare, and manage data for subsequent end user access and analysis, and
- a portfolio of technologies that automate service processes, end user access, and analytical interpretation.

In order to establish a reliable transportation data warehouse, an infrastructure of servers, fiber optic connections, data filtering and analysis, web user interface, video design, and GIS integration is required (Al-Deek et al., 2004).

The main functional and design considerations of data warehouse are data retention and management for the life span of data, appropriate levels of detail, storage capacity and management, data quality, regional data repository and relation to similar national architectures (Turner, 2001). Data protection must also be considered. Once data is stored, it should be protected from unauthorized access or degradation. Software versions for proprietary information should be indicated in a data file (Rasdorf et al., 2001). A data warehouse must balance several potentially conflicting needs. These include providing fine data granularity when needed, providing aggregated data at various levels for data analysis, and providing an acceptable performance for database queries (Smith and Babiceanu, 2004).

A relatively new field for data warehouse is the Digital Library Initiative (DLI). The DLI concerns itself with the automatic creation, organization, and indexing of complex collections of data. There are four main issues of the DLI, they include: 1) central storage of documents, 2) routing documents to interested parties, 3) document version control, and 4) security of documents. Digital libraries have been developed by federal agencies to support the growing need of access to information in a standardized system (Latimer and Hendrickson, 2002).

1.3. Software used in transportation data management

Not so long ago manual data management was the only method available to deal with this huge amount of generated data. This method also suffered from inadequate breadth and depth of metadata (information about the data) (Turner, 2001), and this made the manual management of the data a tedious operation. Technology has responded to the increased demand for the calls to find easier ways to deal with this huge amount of data. It provided sophisticated database management that includes visual and spatial analysis tools (Tate-Glass et al., 1999).

1.3.1 ITS

The most efficient and easiest method to save, access, modify, and share digital data is by using database software. These softwares allow a user(s) to manage data and integrate it to develop reports. The discussion below describes the use of software in both transportation document management and Intelligent Transportation System (ITS).

Many universities have conducted ITS applicable research on data archive solutions. These institutions include Carnegie Mellon, Columbia, Cornell, Stanford, Berkeley, and UC Santa Barbara. In addition to academia, there are commercial data management services offered by, for example, JobDocs.com, Autodesk, and Meridian Systems (Latimer and Hendrickson, 2002).

The 1999 revision of the National ITS Architecture calls for an archived data user service (ADUS) for common ITS data. The tasks of the ADUS include collecting, archiving, managing, and distributing data from ITS sources. The aim is to utilize the ADUS to fuse and format information to produce "data products" for input to federal, state, and local data reporting.

Numerous data management systems have been developed that can be applied to ITS. Two examples of these management systems are the Common Data Format (CDF) and the Performance Management System (PeMS). The CDF is a self-describing data abstraction for storage and manipulation of multidimensional data in a discipline-independent fashion (Kwon et al., 2003). This generic format makes it ideal for ITS applications. The Performance Management System (PeMS) is a low cost and easy to use system that supports the investment decision process. Because PeMS utilizes open software architecture, it can be modified to fulfill specific ITS needs (Turner, 2001).

1.3.2 GIS

Geographic Information Systems (GIS) are specific types of information systems that are suited to the display and manipulation of data with a spatial component. These systems were developed to handle complex data sets for managing and controlling large scale projects (Kimmance et al., 1999). GIS is a generic technology applicable across many disciplines (Goodchild, 2000). The vast majority of transportation analysis and research models contain some form of geographic reference (Goodchild, 2000). In fact, transportation agencies' data are 60 to 80 percent location based (O'Packi and Lewis, 1998).

The application of GIS to transportation systems dates from the earliest beginnings of GIS in the 1960s (Goodchild, 2000), although widespread use is only now being realized. Currently, many transportation agencies use electronically based data management systems for items such as

bridge management, maintenance, planning, traffic volumes, road and lane closures, and infrastructure inventory (She et al., 1999; She, 1997; She and Aouad, 1996). The data required for these management systems all have the potential to be tied to a location and used in Transportation-GIS (T-GIS). In the past, GIS data management has been confined to mainframes, which limited its use. Additionally, the link between transportation databases and transportation models has been missing (Stokes and Marucci, 1995 and Choi and Kim, 1994). As a result, the combined GIS database and transportation model has been underutilized (Guebert, 1991; Quirogo and Bullock, 1996), thus limiting the past usefulness of T-GIS.

Coupling a data warehouse and GIS provides transportation agencies with a valuable tool. Using location as the key data element to link data allows agencies to transform disparate data to useful information. This integration results in a strong tool that can provide historical data, multiple layer support, multiple over lapping routes, updating links and nodes, and synchronization of operational data sets (O'Packi and Lewis, 1998). Additionally, site photographs and videos can greatly improve visual inspection capabilities in the office, digital mapping analysis efforts (Tate-Glass et al., 1999), and bridge management and maintenance (She, 1999). Moreover, the internet can be integrated into these systems to provide easier access to users through search engines and browsers (O'Packi and Lewis, 1998). The Central Florida DOT has such a system. A web based tool utilizing GIS functionalities to collect and distribute real time traffic data (Al-Deek et al., 2004).

A successful T-GIS, produced by the New York City DOT integrated GIS technology with an existing data management system. The developed system is capable of producing special-purpose thematic maps using variables entered by different DOT divisions. Although costly in the design and implementation phases, the integrated system has shown high productivity as well as aiding in management and control of old existing data (Cohn, 1995). Further advancements of the system, such as integrating the GIS with a bridge management system would enable users to easily navigate to pertinent information efficiently. A GIS based bridge management system can provide the following functions: a complete set of bridge information, color coded thematic maps, spatial analysis, online storage and viewing of important bridge events, and useful statistical and visual information for research and design practices. The resultant is a better coordinated bridge inspection program (She, 1999).

1.4. Pros and Cons of electronic document management system

As with any new technology, document management systems are not perfect and therefore have pros and cons. The goal of system selection is to insure that the implemented system has more benefits than costs. Because of the large amount of data produced on a daily basis by DOTs, and the lack of a consistent data archiving and retrieval system, the retrieval process is currently labor intensive and difficult (Quirogo and Bullock, 1996).

Electronic document management systems provide direct access to data, empowering users to make faster and more efficient decisions. These systems enable consistent analysis and provide the mechanism to analyze and predict trends in historical data. Because information such as traffic data, construction and work zone data, traffic incident logs, and traffic control responses is shared, the decision making process can be distributed (O'Packi and Lewis, 1998). Integration

of data across organizational boundaries is a direct benefit of EDM. These boundaries may be internal, such a bureaus or divisions, or external such as DOT contractors (Back and Bell, 1995).

Transportation data comes in many types and formats. For example, the Florida DOT funded research to look at a data warehouse for real time and archived traffic data collected from loop detectors, cameras, and other traffic sensing devices (Al-Deek and Abd-Elrahman, 2002). In the past, DOTs have had major issues in managing the huge amount of data these automated devices collect (Rasdorf et al., 2001). This has become easier and more efficient through the use of Information Technology (IT) and EDM (Bjork, 2001).

In addition to the ease of access associated with digital documents, space savings is also a benefit. Digitally archiving data is highly recommended when storage cost is an issue because a CD can hold 40,000 to 60,000 images while a traditional filing cabinet only holds 14,000 pages. Therefore, approximately four filing cabinets worth of documents can be stored on one CD. Hard copies of scanned documents can then be recycled or destroyed to save storage space. Important documents such as maps should be re-filed after scanning and the location of the actual paper document should be recorded as metadata associated with the digital copy (Rasdorf et al., 2001).

Maintaining the integrity of the database can be difficult for DOTs. Sometimes information is changed without updating the index. Information that is not archived correctly can be lost, damaged, or become inaccessible (Lefchik and Beach, 2006). There is a need to refresh the storage and retrieval technology approximately every 10 years and these electronic data systems require costly regular maintenance and disaster plans. In addition, scanning documents can be tedious and costly, whether done internally or externally. The scanning process can be improved with equipment such as multi-page scanners; however the technology is expensive (Rasdorf et al., 2001). Finally, if the document management system is tied to real-time data it must have the infrastructure in place to handle the high throughput of data. ITS generates large amounts of real-time traffic data, which must be collected, processed, and communicated in real-time (Xiong and Lin, 2000).

The EDM also faces some non-technical obstructions. Implementing such a system requires the collaboration of different levels of an organization (Hajjar and AbouRizk, 2000). In addition, the industry must be ready to adopt the technology (Latimer and Hendrickson, 2002). Employees tend to resist any change; therefore, the management must actively pursue the change and set goals for implementation. The capturing of data from ongoing operation must become part of the daily operations workflow for it to be successful. Treating it as a separate activity will doom it to failure.

1.5. Construction industry document management experience

A relatively similar engineering discipline to transportation that has successfully implemented digital document management is the construction industry. The majority of historic documents stored by state DOTs were generated during the construction of past transportation infrastructure. Because of this, a careful literature review of the construction industry was performed to ascertain the state-of-the-practice with respect to document management. The construction

industry reaps benefits from data archiving systems. In the past, construction data was mostly managed and controlled manually with paper forms and documents being faxed or delivered by hand (Bjork, 2001). Since digital data has become dominant, electronic management and control has become a necessity. The newest information management technologies are being continually adopted by the construction industry. These technologies have the capability to collect, store, retrieve, process, and distribute project data in various formats and from different sources (Caldas et al., 2002).

Document management is a necessity for construction companies. The assigned activities and resources are based upon the information that is available at their sites and offices. Any mistake or loss of data may cause a construction company managerial, financial, or even safety problems (Bjork, 2001). As a result, Electronic Document Management Systems are a must in construction document storage and retrieval. A single building can have 10,000 related documents (Turk et al., 1994). Typical construction EDM has computer representation of the main document body and a reference structure to retrieve the physical document. The EDM must handle drawings, specifications, schedules, quality control reports, and any construction related document. The system can be used to edit, capture, index, distribute, and retrieve various types of electronic and hardcopy documents (Hajjar and AbouRizk, 2000).

The implementation of construction related EDM face some difficulties. The main hurdle in managing construction information is the variety in data types. Data types include structured data files (cost estimating, finance, accounting, data warehouse, etc.), semi-structured data files (HTML, XML and SGML), unstructured text data files (contracts, specifications, reports, catalogs, etc.), unstructured graphic files (2D and 3D drawings), and unstructured multimedia files (pictures, audio and video) (Simoff and Maher, 1998 and Caldas et al., 2002). Standards and software have been developed to deal with the construction industry related document issues.

Another document management tool that was developed for the construction industry is the Construction Document Classification System (CDCS). This system is an environment for the use and simulation of different data selection, data preparation and text classification methods. The system provides flexibility and power for the text classification task and simplifies the classification process for the users. Several methods for data selection, data preparation, and dimensionality reduction were implemented in the CDCS (Caldas et al., 2002).

The construction industry adopted systems from other fields, such as Computer Integrated Manufacturing (CIM), and developed industry specific protocols, such as The Construction Information Classification System (CICS) (Sanvido and Medeiros, 1990). This system creates concept hierarchies that can be used for text classification and can be a standard representation of construction project information. Adopting such a classification system is challenging because of the uniqueness and dynamic nature of each project and those organizations involved (Caldas et al., 2002). There are many additional proposed methods in the literature to manage construction information and some are used to capture, share and reuse project information. Other methods have the ability to extract concepts from textual design documentation. In order to ease access, classification and retrieval of data, the use of arbitrarily structured metadata to markup documents was proposed, in addition to text clustering techniques from heterogeneous

Architecture, Engineering, Construction and Facilities Management (AEC/FM) documents. Further, a controlled vocabulary (thesauri) was proposed to integrate heterogeneous data representations (Fruchter, 1999; Wood, 2000; Brueggemann et al., 2000; Scherer and Reul, 2000; Yang et al., 1998; Kosovac et al., 2000; and Caldas et al., 2002).

A current dilemma facing decision makers in the construction industry is accessing data without interrupting the daily work flow. One proposed solution is the integration of data warehousing and Decision Support Systems (DSS), which would enable the user to identify the right data, locate it, and then provide it to the decision makers so that a decision can be made faster and more efficiently. This integration would result in an improved decision making process (Chau, 2002).

1.6. DOTs experience with document management systems

Several DOTs have started implementing document management systems in their organizations. This has developed due to the massive amount of transportation data and the need to manage and control it on a day to day basis. The following discussion describes the document management systems that are used at example DOTs and their experience dealing document management.

1.6.1 Arizona Department of Transportation

The Arizona Department of Transportation (ADOT) embarked on a document management system by first examining its current document handling procedures. ADOT focused on the gathering and analysis of specific needs and concerns within each of the respective areas of operation within the department. Through interviews and surveys ADOT gained an understanding of the stakeholders' business requirements and expectations (ATRC Research Notes, 2003). ADOT found that the basic challenge was to reconcile the needs of a broad number of business cultures, including ADOT's own agencies, and formulate a single workable strategy.

ADOT's study illustrated a significant need for electronic document management within the organization. The main data source for the system is physical documents currently kept in folders. Electronic preservation and storage of these documents along with reengineered handling procedures increases efficiency and dissemination speed. Departmental borders diminished as a result of electronic document management. The study concluded any transition problems would be due to cultural and not technical issues. The infrastructure exists today to deploy and propagate electronic document management technologies. The study found some legal issues with an electronic system and recommended that these should be resolved before implementation. Finally, ADOT realized the document management market is evolving rapidly and thus recognized the importance of not choosing a proprietary data storage system. In order to stay abreast of new advancements in electronic document management, ADOT joined several industry groups such as the Association for Information and Image Management (AIIM) and the Association of Records Management Administrators (ARMA) (ATRC Research Notes, 2003).

1.6.2 Connecticut Department of Transportation

The Connecticut DOT also has experience with electronic storage methods. In 1985, ConnDOT began utilizing a personal computer and videodisk player to store and retrieve highway images of

the state's 7700 bidirectional-mile highway network. The Photo log Laser Videodisk System (PLV) provided easy access to information about any location on any state-numbered route. This technology proved to be an integral element in the development of the department's pavement management system. Three years later, ConnDOT, in cooperation with FHWA, began investigating the use of the same technologies for the storage and retrieval of bridge related information. The investigation brought to light inefficiencies in the procedures for handling and retrieving bridge data, which had remained virtually unchanged since the department began keeping records. With the Inter-modal Surface Transportation Efficiency Act (ISTEA) of 1991, six management systems were mandated, including one for bridges. At that point, ConnDOT began the development of an information system dedicated to the state's 5000 bridges, the Connecticut Bridge Management Information System (CBMIS) (Lauzon and Sime, 1993). The system was to replace the hard copy binder system used to maintain basic static information about bridges such as route, length, and width along with information from the National Bridge Inventory.

The first action in the development of the CBMIS was to interview department personnel who were familiar with the PLV system and whose duties were bridge related. The goal was to determine what data would be ideally included in the new system. The previous archaic and inefficient system for managing bridge related information caused problems with information sharing and resulted in conflicts when the same work was scheduled to be done by two independent units (Lauzon and Sime, 1993). The new system set out to correct these issues.

1.6.3 Other DOTs

The Texas and Central Florida DOTs have made progress handling real time data. This data is stored for the purpose of analyses, estimations, computer model calibration, congestion monitoring, transportation planning, and pavement design. Florida uses a web based tool with GIS functionalities. The data is a tremendous tool in the hands of the stakeholders and decision makers (Turner, 2002 and Al-Deek et al., 2004).

Virginia DOT is the third largest DOT system in the US. VDOT established a data warehouse as part of its data management infrastructure to improve its productivity (Papiernik et al., 2000). In addition, VDOT successfully implemented a geotechnical data management system for large highway projects. Florida, Kentucky, and Ohio are developing more comprehensive geotechnical data management systems that will provide their users with information about laboratory and in-situ tests, construction control and testing, assets inventory, hazard inventory, maintenance, and research information (Lefchick and Beach, 2006). In fact, FHWA and Ohio DOT formed a group to develop data dictionaries and formats for geotechnical data management systems.

With an increased pressure on DOTs to enhance their productivity, reduce their expenses, and develop new systems, it is becoming vital that DOTs adopt management systems for pavements, bridges, culverts, traffic signs, and other assets (Lefchik and Beach, 2006).

1.7. Conclusions

Transportations systems are complex. They produce huge amounts of data on a daily basis. This data needs a document management system to deal with it. Electronic data management system gives advantages to the industry in storing, processing, accessing and retrieving the different types of data. Data warehouse retains and manages the various levels of data detail, quality and similarity to other data architectures.

The usage of electronic data management in transportation settings eases the problems of dealing with the very large amounts of collected data visually and spatially through ITS and GIS. The coupling of data warehouse with GIS provides agencies with a strong tool that transforms disparate data into useful information. For example, NY City DOT developed a T-GIS system that is capable of producing special purpose thematic maps using variables entered by the different DOT divisions.

Any system has pros and cons. A proper data management system can empower its users with tools to make faster and more efficient decisions. Documents can come in many types and formats, and a document management system can ease the problem of maintaining and providing access to these documents. Finally, a data management system can save space in archiving, but it can also be costly to maintain the integrity of the data management database. Success requires the collaboration and cooperation of the all different levels in the organization.

DOTs such as ADOT, VADOT, ConnDOT, TXDOT and FLDOT, have successfully started using data management systems for their daily activities. Their Data management system helps these DOTs by increasing their productivity and reducing their document handling expenditures through a good and reliable data management system. The remainder of the report focuses on methods for converting, organizing and accessing data from a Document Management System (DMS) will be detailed.

Section 2.0: Electronic Engineering Document Repository Issues

Information Technology makes it possible to create digital repositories of previous hard-copy document archives. The benefit of such a system is to "have data at our fingertips no matter where we are and what documents contains the data." However, such a benefit presumes easy document retrieval. Prototypes of such systems indicate that as point solutions, rapid direct retrieval of documents can be achieved, but as the diversity of documents, user groups, and desired retrieval outcomes increases, the ability to retrieve "the" document and its associated data become less likely. Instead, lists of documents are often found, and navigation through the document list is often cumbersome at best. Business (and even safety) concerns require that existing documents be readily available when needed, that such documents are not lost, and that version and ownership control exists.

Many state departments of transportation currently are undertaking an initiative to scan a wide variety of engineering documents and store them at the folder level. Driven by the inability to locate "the" document has caused severe service disruption and increases safety risks.

The overriding issue is to maintain the integrity of the scanned documents. At the fundamental level, the issue is re-associating pages of the same document once the staples have been removed for the physical process of scanning. The larger problems, however, are the ambiguous language semantics (homonyms and synonyms) relating to project titles, identification schemes, document organization, term redefinition over time, precision, and the absence of robust metadata schematic. This presents a myriad of challenges in establishing appropriate linkages among related but separate documents.

This paper will now examine the problems described in the previous section. In review, these problems can be summarized into three primary topics:

- Re-association of same document pages
- Development of language semantics
- Finding "the" document

Re-association of same document pages are covered in the Organization Process section of this report. The section will provide a roadmap on how to reconnect individual files within a Document Management System (DMS). The process will not only reconnect the files, but also prepare them to be retrieved. Through re-association, the section of different techniques or technologies will be introduced to increase the effectiveness of the process through automation. Though many of the techniques can be accomplished by manual labor, it would more cost-effective to employ the automation techniques under the assumption of a large number of documents.

Once the problem of re-associating has been examined, the Conversion Process section will outline the proper process of converting hard copy files into a DMS. This process will rely heavily on developing proper language semantics and will outline the standards for the actual conversion process. Additionally, the section will describe symptoms of common problems with DMS and where in the conversion process the problem could be avoided. A deeper insight into the techniques utilized in the Organization and Conversion Process sections can be found at the end of the document.

Lastly, once documents have re-associated and new documents are converted into the DMS, the Access section will examine techniques for finding "the" document. The section will discuss various methods of searching, browsing, and previewing documents. Utilizing these techniques will increase the effectiveness of the DMS. Also discussed is how these techniques can be employed together to allow both knowledgeable and novice users the ability to find the document they are looking for.

Section 3.0: Organizational Process

Organization of documents is crucial to the success of a Document Management System. If the steps necessary to properly organize documents are not taken in the initial conversion process from non-electronic files to electronic format, locating and recovering documents is painstakingly difficult.

Unfortunately, disorganized documents are a very common problem. Many companies believe they can save money by archiving their files electronically, but very often select the lowest bidder to make this a reality. The result is millions of single page files, which can be individual documents, or mixed up pages of different documents. All of these scrambled files leave no true way of searching or browsing, making it almost impossible to find the correct document in reasonable time, if at all. This is the problem that is addressed in this section of the paper.

The goal of this process is to take millions of files and organize them in a way that allows them to be searched and retrieved quickly. During this process the following assumptions are made:

- All the hard copy files were destroyed, meaning rescanning is not an option;
- A proper taxonomy was never developed;
- There are multiple document types, such text and graphics;
- Some of the files were originally scanned poorly; and
- There is limited metadata associated with the files.

Top Level Process

The goal of the overall process is to identify and organize millions of single page files. Additionally, the single page files that are parts of larger documents must be regrouped to create whole documents to ensure that all the necessary information associated. The process shown in Figure 3-1 is comprised of two primary tasks that serve to achieve these goals. Sub-processes 1 – 4 defined and establish an organizational structure, while Sub-processes 4 – 6 embed searchable data to recreate whole documents. Once these two tasks have been completed, the documents will be easily retrievable because of the organizational structure.

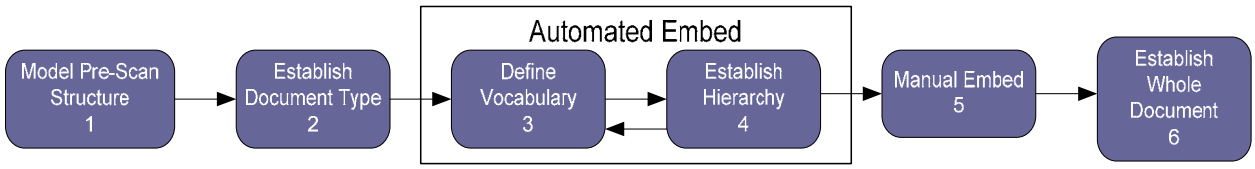

Figure 3-1. Organization top level process map

3.1. Model Pre-Scan Structure

The main purpose of modeling is to accurately model the current taxonomy of the document organization, because the taxonomy defines the structure of the organization and directly affects the development of a hierarchy. The client already has a taxonomy, whether it is formally defined or not. The organization's current taxonomy is used to create the new system. A new system that closely reflects the current system makes transition easier for the users, increases satisfaction with the system and minimizes the reduction in productivity. In order to model the current taxonomy effectively, eight processes must occur, as shown in figure 3-2. The numbering of this section will refer the numbers in figure 3-2 below.

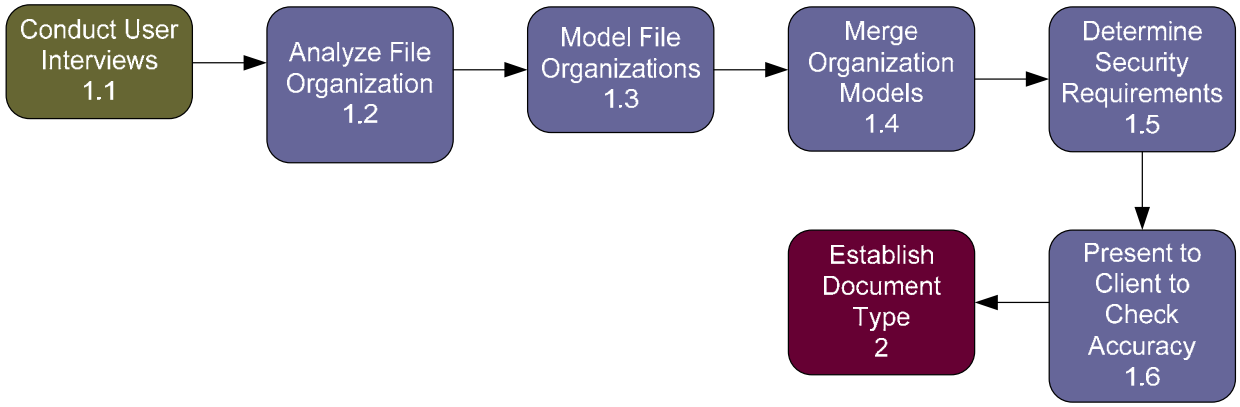

Figure 3-2. Model pre-scan structure process

3.1.1 Conduct User Interviews

The purpose of this step is to gain the user's perspective on how documents are currently organized. This step of the process is important because it helps reveal user's preferences towards the current structure of the organizational system. User preferences offer insight to how the system should remain the same as well as improvements that can be made to the current organizational system.

The wealth of information users can offer about the way documents are organized is crucial to understanding the current taxonomy of the organization. By observing the filing process,

electronic or otherwise, a better understanding of the organizational practices of the company will result. It is important that employees with a lot of hands-on experience with the filing process are consulted for this step to be successful.

User interviews identify the different types of files and how each of the file fits into the organizational structure of the organization. The categorization of different file types reveals the relationships between the various files and the basic hierarchical structure of the filing system.

Interviews produce important insight into how users go about their jobs on a daily basis. Understanding how the users locate documents and determine where documents are filed plays an important role in developing the hierarchy of the system. A hierarchy that matches the current structure of the organization makes the transition from the old system to a new one quicker and easier for users and causes less reduction in productivity. In addition to making the transition easier and minimizing loss of productivity, users will be more satisfied with a system if they are already familiar with the organizational structure.

3.1.2 Analyze File Organizations

The purpose of analyzing the file organization and structure is to find the strengths and weakness of the current system. The interviews allow an in-depth perspective of users' knowledge of the filing structure; however, user perspective may not be completely accurate. Analyzing the structure verifies the information obtained through interviews is accurate and allows for strengths and weaknesses in the current system to be determined.

The file organization can include, but is not limited to, filing cabinets, electronic storage, and documents archives; thus, the knowledge held by any one user may not be completely accurate for the organization as a whole. The information from each interview must be analyzed and combined to find the most accurate representation of the organizational structure. Analyzing the filing structure and the information obtained from interviews is key to developing an accurate taxonomy.

3.1.3 Model File Organizations

The purpose of modeling file organization is to gain a visual representation of the file organization. Models depict the file organizational structure, the taxonomy and the hierarchical structure of the current system. The visual depiction of the file organization helps the client understand the structure of the current organization.

Models of the current structure reveal problem areas and indicate flaws in the organization structure that need to be corrected. The graphical depiction makes the structure more easily understood by others and helps the client understand the flaws in the current structure. Visual representations of the taxonomy of the organization are an important aspect of understanding the file structure of the organization and indicate problem areas within the structure.

3.1.4 Merge Organization Models

The purpose of merging models is to consolidate the ways that the various types of documents are organized and form a complete model of the entire organization system or taxonomy. During the modeling process, many different models are created. The different models represent the alternate methods of file organization (hard-copies, electronic, document archives, etc.). The models have to be combined to illustrate the way in which the organization structures documentation. The overall combination of the models should consider the interviews and analysis done in previous processes.

Merging the models together will indicate flaws or discrepancies between the different organizational practices with different types of documents that were not depicted in models of each individual file organization structure. The overall model is important to depicting how all the documents are organized as a whole.

3.1.5 Determine Security Requirements

It is important to determine security requirements early in the process. The security requirements should detail who has access to which documents and what type of access those individuals have. Additionally, roles should be established for general access to the documents. The roles are important to classify the documents so that each person does not need to be given privileges to access the documents. Instead, a role is given access to the set of documents and each person is assigned to a role. Types of access could include read only, editing, sharing, and administrative.

3.1.6 Client Review

Presenting the models to the client ensures the models accurately depict the current structure of the organization's taxonomy. The client should review the models and provide appropriate feedback and be encouraged to ask questions or comment on anything that does not seem correct concerning the model. This feedback is recorded for use in the next step.

After receiving comments from the users, the necessary changes need to be made to the model. Not all feedback from the client results in a change to the models. The reason the interviews and analysis takes place is to eliminate the possibility of misinformation from the client who might unknowingly be misleading during an interview. If the accuracy of the model is high, it will serve as a better measure for the automated system that is developed later in a project.

Once revisions have been made, the final version needs to be stored so it can be easily located later. When a new system is developed, the model can be used to examine the differences between the new taxonomy and the old one. This also makes the user training process easier, since the differences are on paper and can be explained to the user. Modeling the current system

correctly is crucial to designing a new system that fits the organization's taxonomy; thus, validating the accuracy of the models is a necessary step to ensure quality.

3.2. Establish Document Type

After a current system is modeled, the next step is to begin sorting documents by type (text, image, map, etc.). This is done before determining a company's vocabulary and developing their new file structure hierarchy, because those steps require text-finding tools to search through the digitized documentation. Also, each file's document type must be determined before metadata encoding, because the file type may not warrant an automated embedding procedure. The main goal of this sub-process is that all image files should be pre-screened to determine how their metadata will be embedded so that they may be indexed. There are three possible outcomes of this pre-screening procedure as shown in Figure 3.3., that the sub-process map branches describe. The numbering of this section will refer the number in the figure below. The letters in the figure refers to the techniques in the technique section of this document.

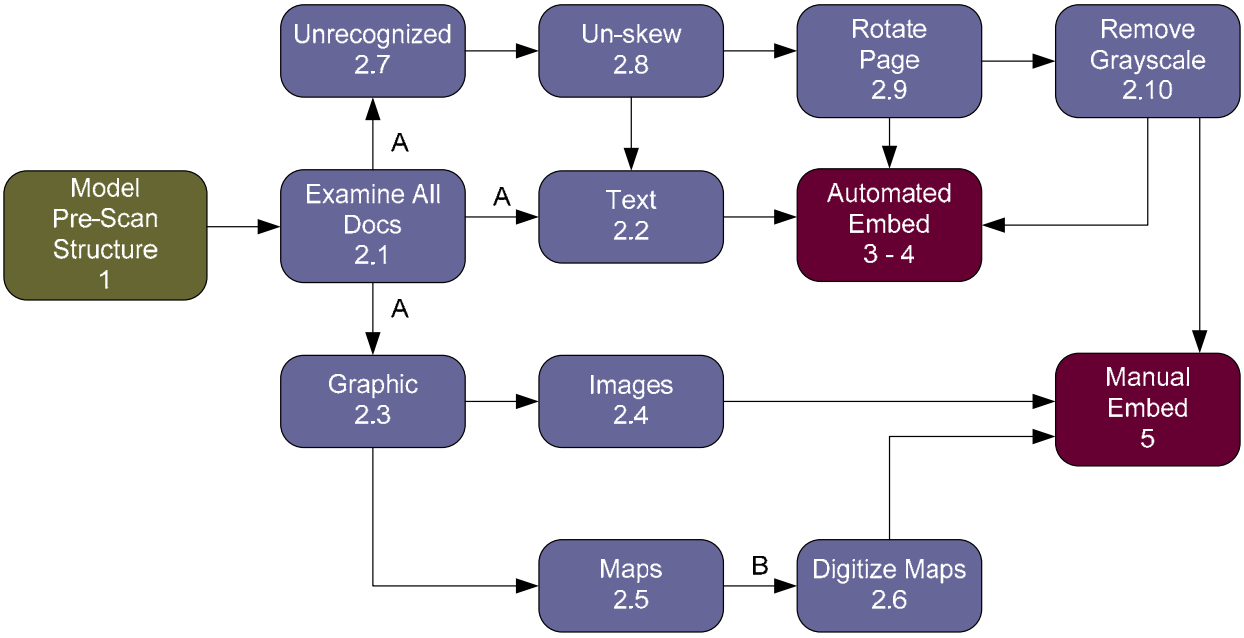

Figure 3-3. Establish document type

3.2.1 Examine All Documents

First run OCR software on all digitized image files. This preliminary run determines basic document types based upon the amount of text found in the file during the OCR examination. Once the type of file is determined, the file can be sent to the correct index embedding sub-process.

3.2.1.A OCR Pre-examination

An OCR program is used to analyze all files that are run through the system. A full image scan must be used here, because the software is looking for any image files that contain text at least a sentence in length. Textually based files will generally contain more than a sentence of characters, and it is assumed that graphical files such as maps may have a few words but will not contain a full sentence.

3.2.2 Text

The best outcome of the preliminary OCR screening is if the file examined was a text file. This is the ideal outcome, because all of the text files gathered can be immediately sent to the automated embedding process for metadata extraction and embedding.

3.2.3 Graphic

Once the documents have been initially scanned with the OCR software, the files that are graphical in nature should be separated out from the other file types. A graphic is any document that is recognized as a non-text item. There is a possibility that bad scans are classified as images; thus, a manual check is done to remove bad scans. Once categorized as a graphic file, these documents are broken down even further into two other sub-categories manually, either images or maps.

3.2.4 Images

Images are classified as any document that contains no text data, an example being an aerial photo. Since no text is present, classification algorithms are ineffective on images. These documents need to be embedded with metadata manually because they contain, in most cases, a completely different metadata hierarchy than text files.

3.2.5 Maps

In this step, all of the maps are differentiated from the other images and are analyzed and sorted once more to re-associate all pages of a single map. This step has to be done manually to ensure accuracy. All maps must be separated and their pages correctly linked before the next phase, digitizing the maps, can begin.

3.2.6 Digitize Maps

After maps have been individually identified and each map has been fully pieced together, the maps should be digitized. This step is optional, especially if the files are to be archived and never edited. However, if a map is going to be used for some purpose other than to be printed, or if the map is a working document and there is a possibility that it needs to be updated in the future, this step needs to be completed. To digitalize maps GIS mapping software is suggested.

Once the scanned maps are downloaded into a GIS, using the on screen digitizing method, they are turned into usable digital data containing x and y (longitude, latitude) coordinates.

3.2.6.B Digitize Maps with GIS

GIS is mapping software that gives the user the ability to convert various types of digitized map images into useable topological data. This is done in by using either a digitizing puck on a digitizing tablet or by on-screen digitizing. On-screen digitizing is the preferred method, because a map that has already been scanned in to the system can be brought up, the user can use the mouse to trace either the topological points (or geo-reference objects) that are depicted on both the current map data or data retrieved from the scanned map (Detwiler, 2002).

3.2.7 Unrecognized

The third possible outcome from the initial OCR examination is that the image is unrecognizable. This process assumes that an unrecognizable file is a badly scanned file. The next couple of processes are in place to attempt to mitigate the effects caused by poor scanning procedures, and thereby create an image that the OCR software can work with. If any of the following solutions make the document recognizable, the document moves into step 2.2 or 2.4.

3.2.8 Un-skew

One possible error from scanning is that of a skewed image. If the image is angled too much, then the OCR software is not be able to recognize any characters in the image. The solution to this problem, of course, is to position the image in such a way that the text is at a ninety degree angle to the page so that the OCR package can recognize characters. Some OCR packages allow template creation which includes the ability to put points in the margin of the page specifically for un-skewing pages scanned in at an angle (Rosen, 2003).

3.2.9 Rotate Pages

Another problem that can be found is a page that has been scanned in upside down. The solution is to simply rotate the page 180 degrees. Some OCR engines will recognize inverted characters, which means rotating the document would not be necessary; however, rotating the document can be coded into software if OCR inverted recognition is available.

3.2.10 Remove Grayscale

The last step in this sub-process handles scans that are of poor quality. Many OCR packages have options that attempt to improve quality of images, such as grayscale removal (improves contrast), despeckling (removes “noise” from an image, including lines or spots), border removal, and background removal (Rosen). If the features of OCR are unsuccessful in identifying the image, the document has to be analyzed for metadata manually. The purpose of attempting to improve image quality is to minimize the number of documents that must be analyzed manually.

3.3. Define Vocabulary

A controlled vocabulary is used to determine the new hierarchical structure of the taxonomy. A controlled vocabulary is a standard list of terms that is maintained to avoid confusion among different words with the same meanings, and is especially helpful in improving search engine results because the words in the vocabulary identify specific document attributes. The model of the current system created in the first sub process can be a great foundation for the controlled vocabulary, and as the vocabulary is refined it can further be improved by using the model in conjunction with an expert in the domain. The taxonomy is the classification technique used by the organization.

This process involves several techniques that extract terms from the text and uses them to classify the document and construct a hierarchy (process 4). In fact, this sub process is heavily intertwined with the creation of the hierarchy. At every level of the hierarchy that is created, this vocabulary is polished. Stated differently, this process defines the vocabulary initially through recognizing words in documents, and refines it with each iteration.

The goal of this sub process is to define a vocabulary that can identify documents as accurately as possible and record the vocabulary into a lexicon or thesaurus for use at a later time. When going through this sub process, the following assumptions are made:

- There is no pre-existing lexicon.
- Documents can be successfully categorized using a controlled vocabulary.

The numbering of this section will refer the numbers in figure 3-4 below. The letters in the figure refer to the techniques in the 'technique' section of this document.

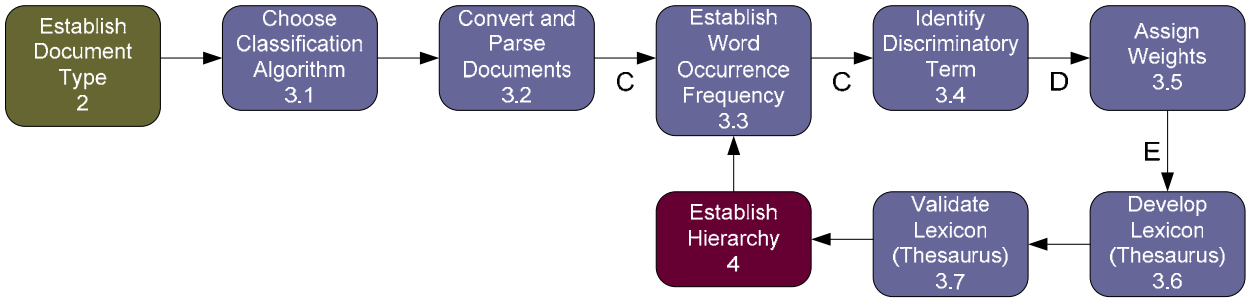

Figure 3-4. Define vocabulary process

3.3.1 Choose Classification Algorithm

The purpose of choosing a classification algorithm is to find the algorithm that best suits the organization. Different algorithms form associations in different ways. Most text classification algorithms use the term "frequency" to establish the importance of words in the documents and use a weighting algorithm to determine the weight of each term identified. Some algorithms,

such as Naïve Bayes, do not take advantage of weighting terms, which can affect the associations identified by the algorithm. In order to affectively choose an algorithm, the advantages and disadvantages of the algorithm are studied and the algorithm that best fits the organization should be chosen. For more information about text classification algorithms refer to technique "C".

3.3.2 Convert and Parse Documents

The purpose of converting and parsing the documents is to separate the terms within each document. The first step is to examine each document individually using OCR. The document is parsed word for word, and the findings are used in the next step to begin the process of identifying that document in the new system. Parsing all the text also allows unwanted tags to be removed. For example, if a file contains a programming language, all the programming tags would be removed to better recognize significant words in the file.

Each word in a document has a different meaning, and some words are more relevant to classifying a document than others. Parsing the document into individual words allows each word to be analyzed for relevance to a documents classification.

3.3.2.C Convert and Parse Documents and Text Classification Algorithms

The technique of text classification algorithms is applied to establish the frequency of every word in every document. After the text in each document is parsed and all unnecessary tags are removed, the algorithm stores all the words for a particular document. The words for each document are stored separately in order identify which words came from each document.

3.3.3 Establish Word Occurrence Frequency

The purpose of establishing the word occurrence frequency is to allow each word in each document and among the documents to be evaluated for relevance to a particular document classification. The more times a word occurs in a document, the more relevant the word is to identifying that document. The more times a word occurs in all documents, the less relevant the word is to discriminating between documents. In other words, if a unique term is used many times throughout a document, but does not appear often in the whole repository of documents, that word is a strong candidate to serve as an identifier for that document. Likewise, if a word is not unique to a specific classification and occurs frequently throughout the whole collection of documents, the word is a poor indicator of the document's classification. Establishing how frequently a word occurs is a key factor in determining the different classification of documents.

3.3.3.C Establish Word Occurrence Frequency and Text Classification Algorithms

The text classification algorithm allows for significant words in each document to be identified. The theory behind identifying significant words is based on the idea that words unique to the domain of the organization that occur in a small portion of documents are relatively good indicators of the document's classification. Words specific to the domain of the organization

that occur frequently in a large portion of the documents are a poor indicator of a document's classification. This technique serves as the bridge between steps, because identifying significant and insignificant words helps to determine the discriminatory words.

3.3.4 Identify Discriminatory Terms

A discriminatory word is a word that falsely attempts to relate two or more documents. The purpose of identifying discriminatory terms is to recognize the words with high frequency throughout all the documents and recognize unique words that still have no relevant subject matter to the classification of a document.

The text classification algorithm identifies word frequencies. Frequency is used to identify and eliminate discriminatory terms. A discriminatory term is a term that poorly identifies documents. Discriminatory terms are eliminated from the vocabulary so that unrelated documents are not accidentally associated. For example, the word "is" would be deemed a discriminatory term because it cannot be used to clearly identify any one type of document.

Enlisting an expert in the field of study or in the organization helps eliminate discriminatory terms. Some terms may appear to be fairly unique and have a low frequency of use throughout all documents as a whole; however, these terms may have no relevance to the field of study.

The discriminatory terms cannot be used to relate documents across the system. If discriminatory terms are not eliminated, documents that possibly have no relation to each other may become associated. Identifying and eliminating discriminatory terms is a crucial step to defining the vocabulary.

3.3.4.D Identify Discriminatory Terms and Weighting Terms

After the discriminatory terms are removed from the vocabulary, the remaining words are significant and can help to identify document relationships or associations. Next, the terms must be weighted. Each terms weight represents its importance. A more heavily weighted word indicates a stronger association of between two or more documents. In the absence of discriminatory words, step 3.4 can assign weights to each word.

3.3.5 Assign Weights

The purpose of assigning weights is to make sure that the most relevant words and are the primary words used to classify the documents. The significant words from each text document are loaded into a vector (one-dimensional list of words, similar to a chain). The words are assigned weights based on the frequency of occurrence on each document and the frequency of occurrence across all documents in the system. If a heavily weighted word indicates an association between documents, an association is likely to exist; however, if a lightly weighted word indicates a relationship, the association between the two documents is probably not as strong.

3.3.5.E Assign Weights and Lexicon Development

Once the text classification technique has been performed and the terms have been weighted, the controlled vocabulary is nearly defined. The control vocabulary is recorded in a lexicon, similar to a thesaurus that pertains only to the specific problem domain. The lexicon is used to provide a reference for identifying the control vocabulary words, the weight of the words, matching synonyms, and discriminating against word homonyms.

3.3.6 Develop Lexicon (Thesaurus)

The purpose of developing a lexicon is to have a recorded list of the control vocabulary. A set of organization key terms must be developed when parsing through documents for organization. Synonyms should be recorded for each of these terms. As more text analysis is conducted through each iteration, more key terms emerge and the lexicon continues to grow in accordance. The lexicon serves as a reference to process 4, which is establishing hierarchy. The process of developing the hierarchy directly relates to the control vocabulary identified in the lexicon and is used to develop association between documents.

3.3.7 Validate Lexicon (Thesaurus)

The validation of the lexicon serves as the quality assurance for this process. Once the thesaurus is developed through iterations of text classification, it must be examined manually. An expert in the organization's field examines the words at the end of each iteration to ensure the proper terminology is being used and recorded.

The evaluation of the lexicon is done after each iteration because the vocabulary becomes increasingly specific as the document associations are formed in process 4. Also, it is important to remember that since this process is iterative, the thesaurus expands and improves with each iteration. As the lexicon is expanded and improved, document classification becomes increasingly specific, which aids the establishment of a hierarchy. Validation of the lexicon helps to provide quality assurance throughout vocabulary definition.

3.4. Establish Hierarchy

After each cycle of defining the vocabulary, a level of the document hierarchy is formed. The hierarchy is largely determined by the taxonomy of the organization. Establishing a hierarchy that fits the current way the organization organizes documents provides an easy transition for users into a new system.

The process of developing vocabulary and establishing a hierarchy is iterative. The vocabulary is defined and is used to establish one level of the hierarchy. After a level of the hierarchy is established, it is evaluated to determine if the hierarchy can be further classified into subcategories (Freeman, 2004). If another level of hierarchy is required, the vocabulary is

redefined, and the established hierarchy process is repeated. If the lowest level has been reached, the folders are developed and the process terminates.

The goal of this sub-process is to develop a hierarchy that closely resembles the taxonomy of the organization and develop the folder structure for the documents. The following assumptions were made when going through this sub-process:

- The defined vocabulary directly correlates with the taxonomy of the organization.
- The hierarchy of the documents is directly related to the taxonomy of the organization.
- The hierarchy of the documents directly determines the folder structure.

The numbering of this section will refer the number in Figure 3.5 below. The letters in the figure refer to the techniques in the "technique" section of this document.

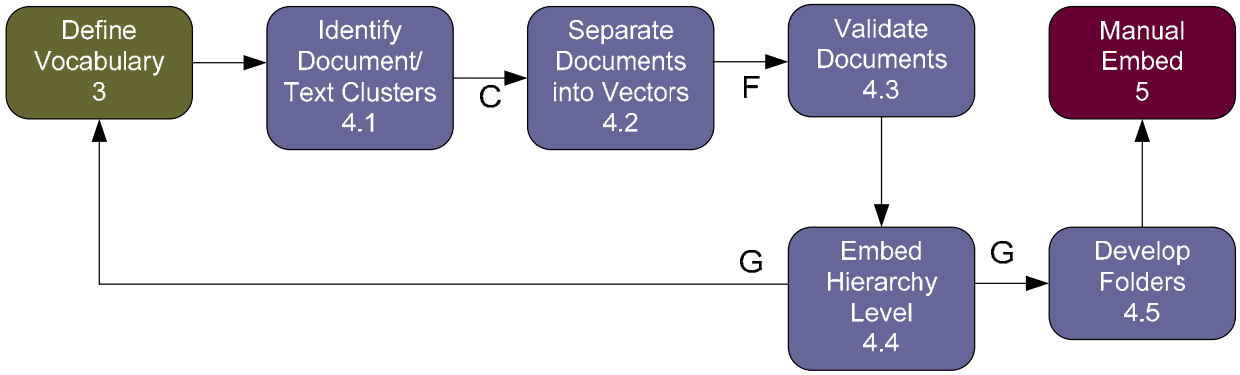

Figure 3-5. Establish hierarchy sub-process map

3.4.1 Identify Document/Text Clusters

The purpose of identifying document/text clusters is to define associations between the documents that are related to one another. The control vocabulary, defined in process 3, is used to identify specific traits contained by a set of documents. Analyzing text clusters containing key terms in the control vocabulary identifies important attributes of a document and allows associations between documents that share similar attributes. The identifying traits shared between documents start broad in the initial iterations of establishing the hierarchy; however, as the control vocabulary becomes more defined, the associations between documents become more specific. Identifying document/text clusters allows associations between related documents to be created (Munteanu, 2005). If the text classification algorithm cannot identify the document, then it has no associations and is not able to fit into the hierarchy. These documents must be sent to the Manual Embed sub-process.

3.4.1.C Document/Text Clusters and Text Classification Algorithm

The technique of text classification is used in identifying text clusters. Text classification uses algorithms and weighting techniques to identify significant words and discriminate words. Significant words carry a lot of weight and documents are associated through the similar uses of these significant words (Freeman, 2004). The significant words and weights were determined and recorded in process 3.

The discriminate words are words that can falsely associate documents together. Control and discriminate words are derived by the frequency of occurrence in documents and by experts in the field of study (Freeman, 2004). For more detail on frequency and weighting of control and discriminate words, refer to the technique index sections C and E. The benefit of using text classification algorithms when clustering documents is that the defined vocabulary associates documents together through the use of text patterns (Freeman, 2004). As the hierarchy becomes larger and more defined, the specific associations between different text clusters become more defined and specific, allowing more closely related documents to be associated.

3.4.2 Separate Documents into Vectors

The purpose of separating the documents into a vector or chain is to group all associated documents together into nodes, while maintaining the relationship of all the documents. A vector, in this case, is an array of document clusters ordered so that an individual cluster can be located with a single index. For example, if the documents of an organization are clustered by department, department would be the index and a cluster can be located by the department to which they belong. A node is a joint in the vector. In the example above, each node is a cluster of documents belonging to a department of an organization.

In the initial iteration of establishing the hierarchy, there could be a significant number of documents that are seemingly unrelated in the same node. In the initial iterations of both defining vocabulary and establishing the hierarchy, the control terms are more generalized (Freeman, 2004). For example, if each node represents the documents relating to a department in an organization, the node may contain various types of documents. The relationship between these documents is the department in which they belong; thus, they are related even though they may contain completely different subject matter. As the vocabulary becomes more defined, the relationship between documents becomes more refined. Refining the vocabulary allows for smaller document clusters with more closely related subject matter. Separating the documents into vectors allows clusters of related documents to be grouped together.

3.4.2.F Separating Documents into Vectors and Tree Structure

The use of topological tree structures in separating the documents provides an easily understandable hierarchy. In this case, topological refers to the collection of sets of documents. Once the documents associations are defined, the tree structure must be formed. The text

classification algorithm provides association between documents; however, not all documents share the same associations (Freeman, 2004). Thus, each cluster of associated documents forms a branch as shown in Figure 3-6.

All associated documents are stored in a node and each node is stored in a vector or chain. Storing the documents in a vector makes linking documents together quick and easy. The root of the tree is made up of a node consisting of all documents; the level one hierarchy consists of a vector containing multiple nodes. Each child node in the level one consists of associated documents. The nodes in level one are considered child nodes because they broke off from the root node or parent node. Each node in the level one can be separated into more child notes to create a larger hierarchy. Each child node remains stored in a vector to ensure that the original associations in the parent nodes remain consistent (Freeman, 2004). (Refer to technique index section F for more information.) Using this tree structure technique, documents are separated into classifications while maintaining their original associations with other documents.

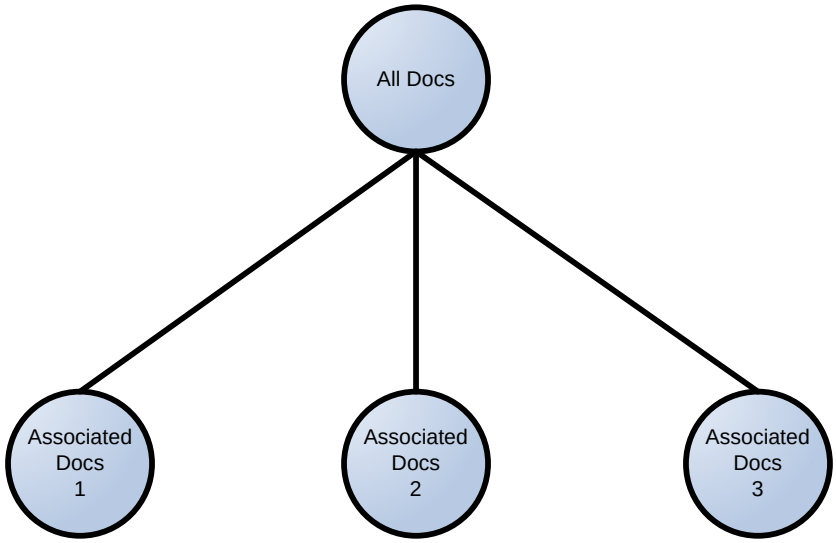

Figure 3-6. Associated document tree branches

3.4.3 Validate Documents

The purpose of validating the documents is to avoid the association of unrelated documents into clusters. Validation allows for quality assurance and should be done during all iterations of establishing the hierarchy. Validating the association between documents is extremely important during the initial iterations of establishing a hierarchy because the control vocabulary is very general, and incorrect associations are more likely to occur.

To keep the validation process more simplistic and less labor intensive, a set of training documents are used initially. Training documents are a set of sample documents that span every

type of document and are representative of the entire population of documents. The training documents allow statistical measures to be used to determine the accuracy of the control vocabulary and document clustering algorithm (Halkidi, 2001). If the training documents are representative of the whole population, the percentage accuracy found using the training documents closely correlates with the accuracy of the whole population.

If the percentage of accuracy found when using the training documents is not high enough, modifications should be made to the accuracy percentage. To increase accuracy, the redefinition of vocabulary at a given level of the hierarchy may be necessary, or alterations to the classification and weighting algorithm should be made (Caldas, 2002).

During the validation, the vocabulary used to associate documents should be the main determinate in whether the documents are closely related. For example, if documents are being associated by department, the vocabulary terms that indicate the department to which a document belongs should be the primary deciding factor to whether to document association is correct. The use of validation provides a method of quality assurances and allows corrective action to be taken if document clusters are not accurate.

3.4.4 Embed Hierarchy

The purpose of embedding the hierarchy is to identify the hierarchical level of each document, to allow for the development of a folder structure, and to place the documents into the appropriate location in a later process. After the validation of the document association, the level of the hierarchy is embedded into each document within the node. As the process of establishing the hierarchy continues, each level of the hierarchy is embedded. By embedding each hierarchical level, navigation to each documents location in the hierarchy structure is possible (Freeman, 2004). Once the lowest level of the hierarchy is reached and embedded into the documents, all documents are organized according to the defined vocabulary and the organization taxonomy.

3.4.4.G Embed Hierarchy and XMP

The embed hierarchy process takes the information provided by the validation process and embeds the hierarchy metadata into each document within the nodes. The method of doing this is called XMP.

The embed process takes the hierarchical metadata and uses a XMP technique called metadata streaming. The metadata stream essentially forms an information dictionary for the hierarchy structure. The metadata is stored in XML packets and attached to each document (Embedding, 2001). The embedding of such metadata allows the recording of the hierarchy. For more detailed information on XMP, refer to technique index section G.

3.4.5 Develop Folders

After the hierarchy has been determined and embedded in each individual document, a folder structure to store these documents must be created. The purpose of developing a folder structure is to establish a physical representation of the document hierarchy, which is a representation of the organizations taxonomy. In order to develop the folder structure, the root folder is created. The root folder contains all subfolders, which are derived from the hierarchy. After the root is created, the first level of the hierarchy is obtained from the embedded hierarchy in documents. Next, each subsequent level of the hierarchy is obtained and the folders created. This process continues until folders have been created for each level of the hierarchy. Refer to Figure 3-7.

At each level of the hierarchy, the naming convention of the folders is determined by the hierarchy level. The naming convention is predefined by the organization according to the taxonomy and the subject matter of the level in the hierarchy. The naming convention also correlates with the classification or category of the document the folder contains. After the lowest level of the hierarchy is reached and the folders have been created, the physical hierarchy of the system is represented by the folder structure.

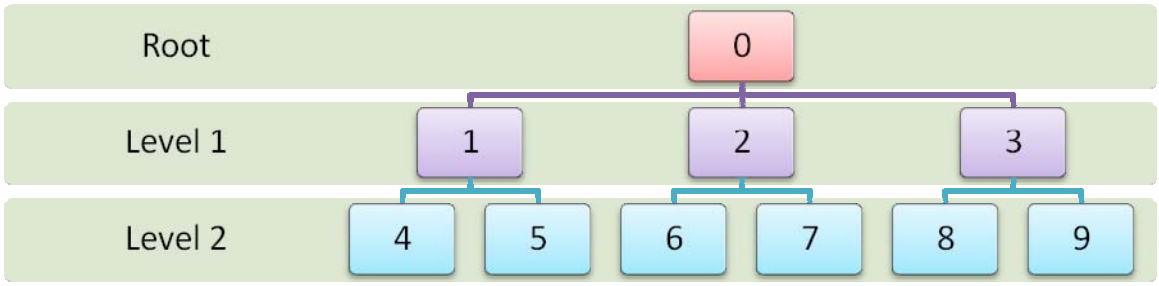

Figure 3-7. Standardized folder structure

3.5. Manual Embed

The purpose of this process is to take the documents that were not recognized by the OCR software and categorize them, thereafter embedding them with proper metadata. This sub-process takes place after the hierarchy has been established, because before the metadata embedding can begin, the metadata categories to be embedded need to be defined.

Metadata is a nametag for a document; essentially it is information about a document that is stored as a part of the document. Defining the metadata is required first, because if random bits of metadata are recorded into each document before a standard is created, there is no way to know how to access these documents. This affects the entire embedding process which then needs to be redone.

Each document is then screened for information relating it to any of the pre-determined categories. Once discerning information about the file has been pulled, this data must be stored in the file for indexing purposes.

To verify accuracy of process, tests need to be performed on the document. Depending upon the outcome of the verification, the document is either moved on to the Establish Single File sub-process, or a failure would result in the document being sent back through the embedding sub-process.

The numbering of this section will refer the number in the figure below. The letters in the figure refer to the techniques in the technique section of this document.

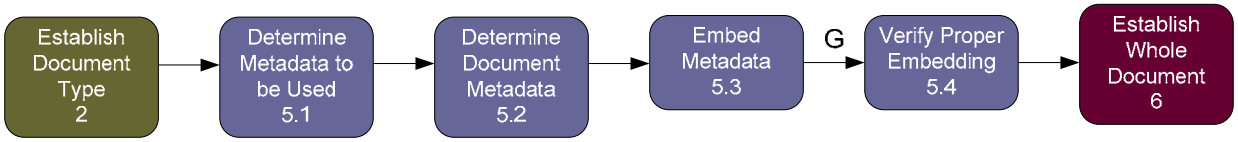

Figure 3-8. manual embed sub-process map

3.5.1 Determine Metadata to be Used

The first step of manually embedding the metadata is to determine the metadata to be used. This step should be done utilizing the defined vocabulary and hierarchy created in sub-processes 3 and 4. Each layer of the hierarchy should be a metadata tag. The defined vocabulary and hierarchy is directly related to the organization's taxonomy; thus, using the vocabulary and hierarchy to determine the metadata used gives the manual embed sub-process structured guidelines for determining the metadata. Determining the metadata first allows a better diagnosis of the data for each document.

3.5.2 Determine Document Metadata

Once the metadata naming scheme has been developed, the document is searched for content linking it to each of the metadata categories in the previously chosen step. A person must to read the document in order to discover material that identifies the document and determine which metadata categories apply. The specific words and phrases identified during the vocabulary defining process should particularly be searched for, as they almost always help identify the document. The goals of this step are to successfully find all pertinent metadata contained in the document text and, to ensure that this file is properly stored and is easily accessible.

3.5.3 Embedding Metadata

After the files have been searched for their appropriate metadata, this information must be stored within each file. The purpose for this is to have all information that identifies and categorizes a

particular file available. Files with metadata embedded within them are more instantly recognized by search tools, and already contain secondary indexes that associate them to similar files.

3.5.3.G Using XMP to Embed XML Metadata

XMP standards not only state the best information to store as metadata, but they also explain the best way to embed metadata into files. This is generally done by placing the information into an XML packet or metadata stream. According to the standards, the packet is attached to the end of the file, but the stream is fused inside the file. This metadata package is created either with software designed for that purpose, or it can be manually coded (Adobe, 2001).

3.5.4 Verify Proper Embedding

Since computer related processes are never one hundred percent accurate, this step, which verifies that the metadata was embedded properly, is needed. There are two methods for validating accuracy. The first is to open the document within a code editing/viewing program. This can actually be done using Microsoft's Notepad application. After the file is opened a person must manually search for errors. The second option is to place the file into a Document Management System and perform a search to test for each group of the embedded terms.

3.6. Establish Whole Document

Assuming that all paper documents were scanned into the system one page at a time, and each page digitized in this manner has its own separate electronic file, all of these single page files must now be combined to create the document that they make up. Since all of the single page files had their hierarchal metadata embedded in the previous sub-process, they are now ready to be combined into a full document.

The numbering of this section will refer the number in the figure below. The letters in the figure refer to the techniques in the technique section of this document.

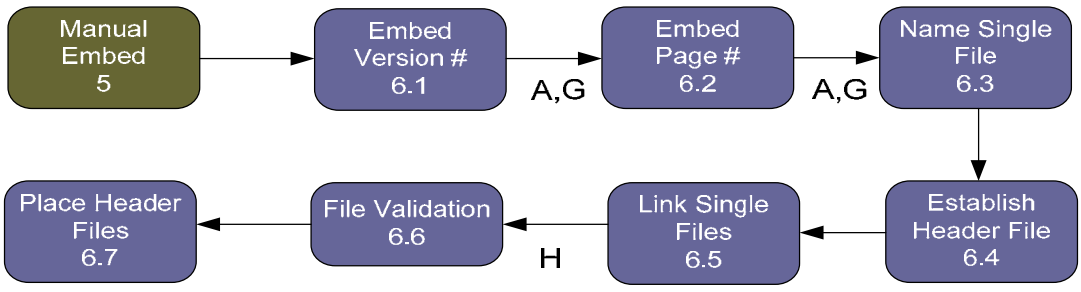

Figure 3-9. Establish whole document sub-process map

3.6.1 Embed Version Number

Once all hierarchical metadata has been encoded into the files, but we must still establish which version of the document this file represents must be established. This must be done before determining page number because different versions of the same document are not identical. OCR software runs through the document to determine the version number of the document page in question. Searching the entire document may not be necessary. If version numbers are typically present on a certain area of a page, only that particular area should be searched in order to increase the efficiency. Once the version is determined, XMP is used to embed the data into the file.

3.6.1.AG Searching with OCR and Embedding with XMP

Typically, document versioning follows a specific format and is located on a particular area on the page. For this reason, the OCR software can utilize zone search to limit its scope to only include the area where the version number is expected. This shortens the time it takes the software to find the versioning information. A template (search for #.# formatting) can be used in conjunction with zone search to further expedite the search.

The version of the document being analyzed is stored in a XML packet. The packet is combined with the file using XMP metadata embedding standards. According to the standards, this packet can be attached to the end of the file, or fused inside the file. This packet is created either with software designed for that purpose, or it can be manually coded. For more information about how this is done see Technique G.

3.6.2 Embed Page Number

After the version number of the single file is embedded, the last piece of metadata to find and embed is page number. The order of each file in the full document cannot be determined unless its page number is found. OCR software searches the file for the page number, and once it is found the page number is embedded into the file using XMP. This step is the last embedding process, and as such completing this step allows the task of combining the single files into a full document to begin.

3.6.2.AG Searching with OCR and Embedding with XMP

This technique is used in the same manner as searching for the document versioning with OCR and embedding versioning metadata. See 6.1.AG.

3.6.3 Name Single File

All single page files have been embedded with all metadata associated with them. Now, all of these files need to be renamed so that in the event that they need to be accessed, they can be found. A standardized naming convention should be decided upon before beginning to rename these files; a good convention to use would be that of the hierarchical metadata.

That would mean long file names, but they would be very precise. Naming the files should be done before linking the files into full documents, as changing file names afterwards could corrupt the links.

3.6.4 Establish Header File

A header file for each document is now created. The header file is the cornerstone that makes the Link Single Files step of this sub-process possible, because the header file holds the links to all of the single page documents. This header file is later renamed to reflect the full document it represents. Once again, a naming convention should be created before beginning to name the files, and the header file names should not be the long string of hierarchical metadata. In other words, the name should be short and specific to the document it embodies.

3.6.5 Link Single Files

After the header file for each document has been created, all of the single page files are now combined into the full documents they are part of. Instead of combining each individual file into a whole document, the best way to combine these individual files is to link them all together. In this way the files are associated without having to actually merge them. This is beneficial in two ways. One, there would not have to be a lengthy process undergone to join all of the single page documents. The second is that should files be wrongly associated, they would be much easier to correct if linked in such a way.

Since the header file is the only file that is accessed by users, the header file must now inherit the metadata from the single files associated to it. The metadata is not specific to the single pages, but describes the attributes of the full document itself. This metadata provides both a means for the header file to be accessed by search engines, as well as a way for the header file to be placed correctly within the hierarchical folder structure.

3.6.5.H Virtual Association Links Files

Virtual Association is a conceptual technique to describe the way the individual files are linked. The header file stores links or pointers to all of the single files. When the header file is accessed all links are opened and the single files are all combined dynamically into a viewer. This viewer would be a read-only, single run instance that is discarded when the header is closed.

3.6.6 File Validation

File validation is performed when the file is first accessed, because a step to manually verify that each file is error free would take up an unreasonable amount of time. This is a quality assurance step to ensure that the file is properly linked; that is, that the document contains all of its pages in the proper order, and none of the pages are from a different document.

A master list of all documents is created which is then updated when users access each document. Every document has a visual queue when accessed denoting its status on the master

list. Documents that have been determined correct are highlighted by a checkmark icon beside the document name. If some of the links are to the wrong file or the document is missing pages, then the problem should be written out in the master list by the user who accessed the file. Users should also note correct files for submission to the master list.

An expert or administrator must periodically access the master list to resolve problems in documents noted by users. If a document has an incorrect link in its association, the problem can be fixed by simply deleting the link to the incorrect file. A document that is missing pages must be marked incomplete on the master list so that users accessing the document are aware. One way the administrator should attempt to identify pages that are missing from a document is to look at the time it was created. Since pages of a whole document were most likely scanned at once, there should be a level of accuracy to looking at pages that were scanned in a very short period of time.

3.6.7 Place Header Files

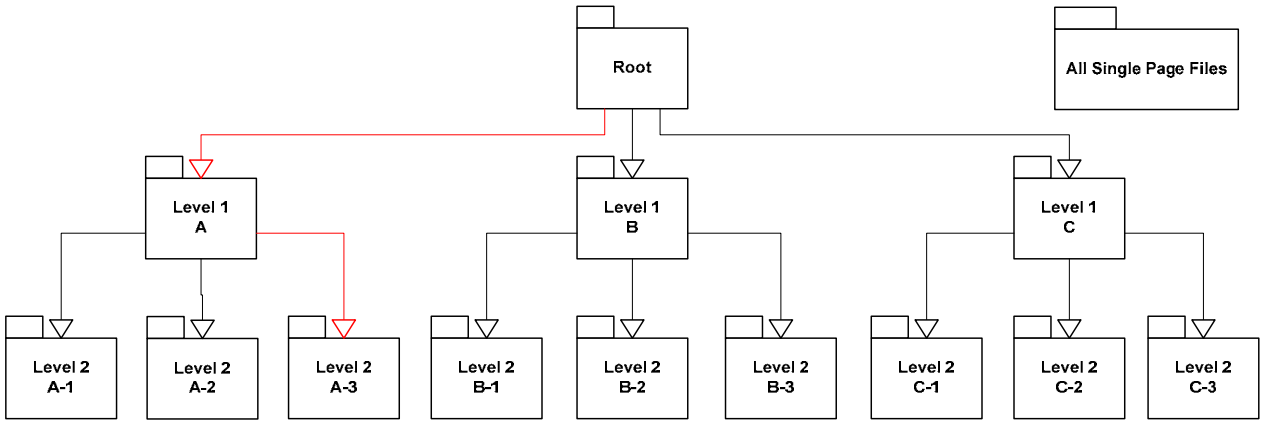

Figure 3-10. Placing documents using folder structure

After all files have been associated, the final step is to place the header files into their appropriate folders. This step greatly resembles the "create folder" structure step in the Establish Hierarchy sub-process. The highest Hierarchy Level, or root, is where all files are placed at first. The header files' metadata elements are then compared to this hierarchy level's entire collection of sub folders (the next level down). The files that match lower level folders are physically moved down into that folder level. An example is shown in figure 3-10. The header file is moved from the root folder to level 1-A, then it is categorized further down to level 2-A-3. This step is repeated for all documents until all files reach their corresponding lowest level folder.

Note that only header files are placed into the folder hierarchy. The header file is the only file being accessed by end users of the system, therefore the single page files that make up the whole document need not be inside the folder hierarchy. The single page files are all placed in the

same directory. They are only accessed by the header file's pointers when users open the whole document in the viewer.

Section 4.0: Conversion Process

Document conversion is the foundation of every Document Management System (DMS). If the work that is conducted while scanning documents into the system is careless, the robustness or capabilities of the DMS are drastically reduced. This section of the paper focuses on outlining the best practices used for entering documents into a DMS. Additionally, the most common mistakes are examined, particularly where they typically occur in the correct process of conversion from paper to electronic format.

The process presented below applies to two situations:

- Converting documents into a completely new system with no form of organization.
- Converting documents into a system with established rules of organization that are working properly.

Assumptions made for this section:

- The DMS is centralized.
- The DMS is for archival use.
- Documents are currently organized (except for completely new companies).

Top Level Process

The goal of the conversion process is to convert documents into an electronic DMS as accurately as possible. The first step defines the controlled vocabulary of terms that are used to structure and reference the files. This step is the foundation for the whole system and if not done properly highly reduces the likelihood of an effective system. Document Preparation addresses the specific needs that could arise for any given document. The third step, Digitize Document, ensures that the conversion of the paper copy into digital format is done as efficiently as possible.

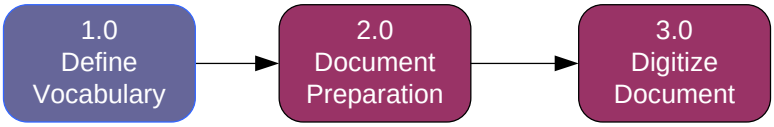

Figure 4-1. Conversion top process map

4.1. Define Vocabulary

A controlled vocabulary is an organized listing of words that have been designated as appropriate for referencing files. The vocabulary serves as a means to reduce the gap between the words a user might enter to search for a document and the words that are actually attached to the

document, therefore increasing the ability to retrieve the document. The vocabulary must be continuously built upon and improved to ensure accuracy of search results within a DMS.

The steps for this process are necessary for new systems with no organizational structure, and explain what needs to be done to successfully develop an effective controlled vocabulary. The iteration within the process refines the vocabulary with each loop, increasing the number of synonyms for each word in the vocabulary that a user would likely use to search the documents in the system. The numbering of this section will refer the number in the figure below.

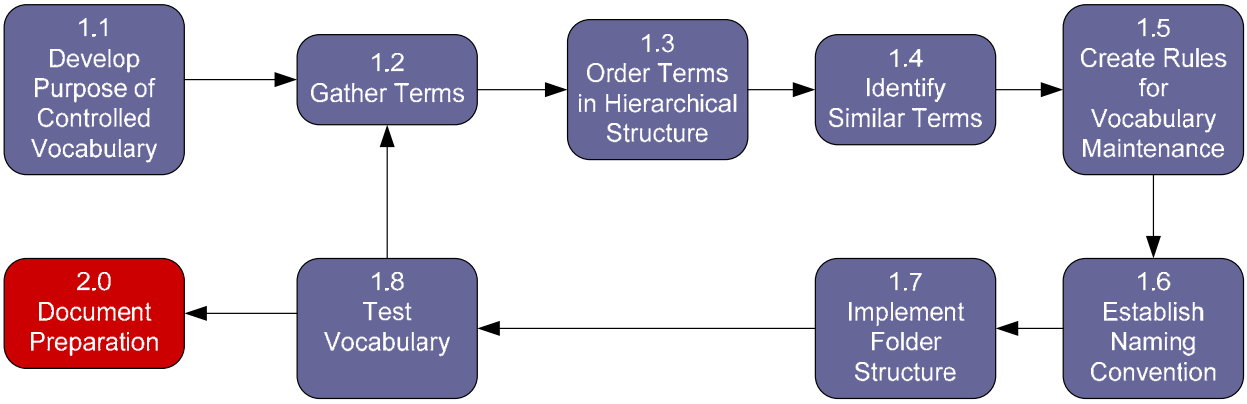

Figure 4-2. Define vocabulary sub-process

4.1.1 Establish Purpose of Controlled Vocabulary

The first step in creating a controlled vocabulary is to decide exactly what the purpose of it is going to be. The terms within the vocabulary can be used to improve searching, browsing, or both. With searching, a strong thesaurus needs to be developed in order to define all synonyms, homonyms, and common misspellings of words in the vocabulary. For example, if the word "bow" is used in a search, then files containing the word referring to a bow and arrow, the bow of a ship, or a bow-tie all need to be returned, since the intention of the user is not exactly clear. This would be an example of the need for homonyms in a controlled vocabulary. This is further explained in step four of this process (Fast, Leise and Steckel 2003).

It is important to establish how specific every term is in relation to the others to improve browsing capability with the vocabulary. The details of how to accomplish this are detailed in step three of the process.

4.1.2 Gather Terms

The goal of gathering terms is to find the words that are the most similar to what the end user will search for. For many industries, vocabularies have already been developed and are available for purchase. If a vocabulary is available, much time can be saved by refining a pre-built vocabulary. User interviews are useful in determining what terms from pre-built vocabularies are used the most by employees.

If no pre-built vocabularies exist, then the documents must be examined and the key terms that can identify documents need to be added to the vocabulary. Since defining a controlled vocabulary is an iterative process, a beneficial way to continue building it is to track the words users search with, measure the success rate of those terms, and add the appropriate ones to the vocabulary (Fast, Leise and Steckel, 2003).

4.1.3 Order Terms in Hierarchical Structure

Once the list of terms is compiled, they must be differentiated based on how specific they are. Related terms should be grouped together. Throughout the vocabulary, decide which terms are broader and which are more specific. Essentially, an organizational structure resembling a pyramid is created, with the most generic terms on top and the most specific on the bottom, eventually leading to individual files.

Unless the system under review is for a completely new company, it is necessary to examine the current folder structure that is used. This assists in determining where any major changes are made from the organization the users are familiar with, and therefore make it easier to explain the changes to them.

4.1.4 Identify Similar Terms

This process involves building a thesaurus. Synonyms and homonyms must be identified so that if a user meant to search for something and actually used a slightly different term, the correct results are still returned. If the term “baby” is used for searching, synonym results should also be returned containing the terms toddler, infant, etc. If the term “hare” is used for searching, the homonym results for hare and hair must be returned. Common misspellings must also be identified. To demonstrate the importance of having common misspellings included in the vocabulary, a search containing misspelled words on one of the many online search engines typically displays a result saying “Did you mean”, followed by the correct spelling. An excellent vocabulary will have common misspellings built in (Fast, Leise and Steckel, 2003).

4.1.5 Create Rules for Vocabulary Maintenance

Vocabularies are measurable by their effectiveness to produce relevant search results. This can become compromised by a number of different things, but most often is due to the fact that

vocabularies can become dated if not properly maintained, especially if the terminology within the vocabulary is constantly changing.

When decisions are made as to where to place terms within the hierarchical structure, record them along with the reason for placing them there. Additionally, the organization creates a schedule for how often the vocabulary is updated. Establishing precedents and rules for new terms helps with adding new terms into the vocabulary later, especially if it is maintained by different people over time. This also increases consistency and reduce the learning curve for the vocabulary (Fast, Leise and Steckel, 2003).

4.1.6 Establish Naming Convention

In order to name the files, a naming convention should be established utilizing the vocabulary and hierarchy. To transition easily from paper files to digital documents, the naming convention chosen needs to be a very close representative of the current naming convention in place. The benefits of using a naming convention include:

- Allowing the sorting of documents in logical sequence.
- Helping users to search for documents and identify the items they are searching for easier.
- Tracking versions.

Some of the most common ways to name an electronic document are title, version, number, date of creation or newest version date, author or creator, type and file extension. If the chosen DMS is able to use and search metadata, then shorter names can be used such as title, version number, and date only (Alberta Government, 2005).

4.1.7 Implement Folder Structure

The purpose of implementing a folder structure is to establish a physical representation of the document hierarchy. In order to develop the folder structure, the root folder is created. The root folder contains all subfolders, which are derived from the hierarchy. After the root is created, the first level of the hierarchy is developed from step 1.3. Next, each subsequent level of the hierarchy is obtained and the folders are created. This process continues until folders have been created for each level of the hierarchy.

4.1.8 Test Vocabulary

Testing the vocabulary is a process that occurs after the document system is populated with a sufficient number of documents. Search with terms in the vocabulary that return a known set of documents in the system. If this is successful, then the vocabulary is already demonstrating its usefulness. After using those terms, search with synonyms of the same terms. Similar results should be returned if the thesaurus was built properly. If the search does not return any relative

documents, or returns an unsatisfactory number of relative documents, the process needs to iterate back to step 1.2 (Fast, Leise and Steckel, 2003).

4.2. Document Preparation

The steps outlined in the document preparation sub-process are heavily based on the Library of Congress standards for digital conversion of text and graphic materials. This process is the quality control step for the top process, and ensures that all of the precautions and preparations are taken prior to document conversion that results in the highest quality of digital reproduction. The numbering of this section will refers the number in the figure below.

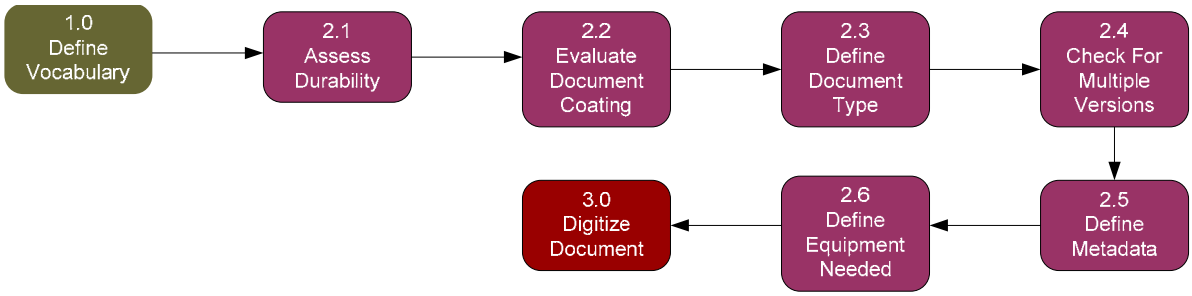

Figure 4-3. Document preparation sub-process

4.2.1 Assess Durability

Scanning a text or image material into a digital format should never compromise the condition of the original document. Materials that appear old or worn need a conservation assessment prior to scanning. If necessary, a conservation treatment or re-housing is applied to the document before scanning.

Materials to be scanned that are bound might require book cradles to support the spine during the scanning. Documents that have been folded for a long period of time tend to tear at the fold. If at all possible, avoid scanning these documents with the crease pressed flat against the scanning bed. The part of the document that is not being scanned needs to be supported to prevent damage to the crease.

4.2.2 Evaluate Document Coating

Before scanning a document, the protective coating must be taken into consideration. If the document is laminated, the glare produced during the scan could result in poor quality. Transparent protective sleeves could produce a similar result. If the integrity of the document is not compromised, remove it from the sleeve before scanning it. If the document is laminated,

several sample scans might have to be performed until a configuration is found that eliminates the glare.

4.2.3 Define Document Type

Different types of documents are held to different standards. This includes whether the document is text, image, picture, etc., and whether or not OCR is used on the document. The standard resolution for all documents is 400 pixels per inch (ppi) in the center of the document with minimal loss of quality in the corners. However, for documents not being scanned for OCR purposes, a resolution of 300 ppi is considered acceptable only if due to device or format limitations. Additionally, OCR documents need to be scanned with a minimum of an 8-bit grayscale bit depth. If color is an important attribute to the document, then the bit depth needs to increase to at least 24-bit color.

Maps have slightly different standards. If maps are to be used for content research only, the standard minimum is 250 ppi. If the expected use of the map is for reproduction, the quality needs to be 400 ppi. It is important to note that the ppi may need to be adjusted if the file size approaches 500 megabytes, especially if the map pieces are separated and must be re-attached. Bit depths of maps must be 24-bit color at a minimum. Additionally, because many maps only have a text scale which does not translate on a computer screen, scale conversion data should be added to maps.

Photographs and graphic arts have separate standards. If access to content is the only expected use after conversion, the resolution minimum is 300 ppi with 8-bit grayscale, unless color is important. If document reproduction is the expected use for the file, the resolution needs to be set to the device's maximum setting with 24-bit color minimum.

If a report or another type of multi-page document is being converted, it is imperative to define the complete document and convert it so that the separate pages of the document are associated. Executing step 2.5 correctly accomplishes this.

4.2.4 Check for Multiple Versions

In many organizations, documents can be modified in several different locations. This results in different versions of the same document, and can lead to confusion over which version is the most recent or correct. When this is the case, there are several available solutions.

Virtual association can be used to attach the different versions of the document together, so that the user can browse for the version they are looking for. Each document would contain links to the rest of the versions of the document. Another option would be to convert the different documents into the system as different versions of the same file. This method would also provide the users with a choice of which version to retrieve. If the above options are not

available, converting all of the different versions of the documents and storing them in one folder in the system is a possibility.

4.2.5 Define Metadata

Each document should have specific information that defines it, such as the folder that contains it or the filing cabinet where has been stored. A majority of this information is obtained from the hierarchical structure created in step 1.3. This data about the document is identified and recorded so that it can be embedded into the document. It is important that an employee that is familiar with the content of the documents has the responsibility of defining a document's metadata.

4.2.6 Define Equipment Needed

A scanner is acquired based on the standards defined in step three. If the system is for storing text documents only, then a lower quality scanner can be used. If photographs are stored in the system, a scanner capable of high ppi scans is necessary.

Scanning maps requires special equipment due to their size. A0 scanners are considered standard for scanning maps, with scan width options ranging from 25" to 54". While a 36" to 42" scanner is sufficient for most maps, a 54" model provides the most comprehensive solution.

If a large document is being scanned, it is easier to lay it on a flat surface than to feed it through a scanner. Also, if there is any concern as to the structural integrity of a document, a flatbed scanner has a much lower chance of damaging it.

4.3. Digitize Document

In this step the paper document is scanned into the system and a digital replica is made. To make this possible the scanner must be prepared by running test scans and ensuring that the proper equipment is being used as discussed in step 2.6. Once the scanner has been prepared, a header file is created. The header file contains all of the metadata for the document that it makes up. As each of the pages of the document is being scanned, a pointer from the header file to the newly scanned file is created. After the whole document has been scanned into the system a single document is established. Finally a quality assurance step is taken to ensure that a document meeting all standards has been created.

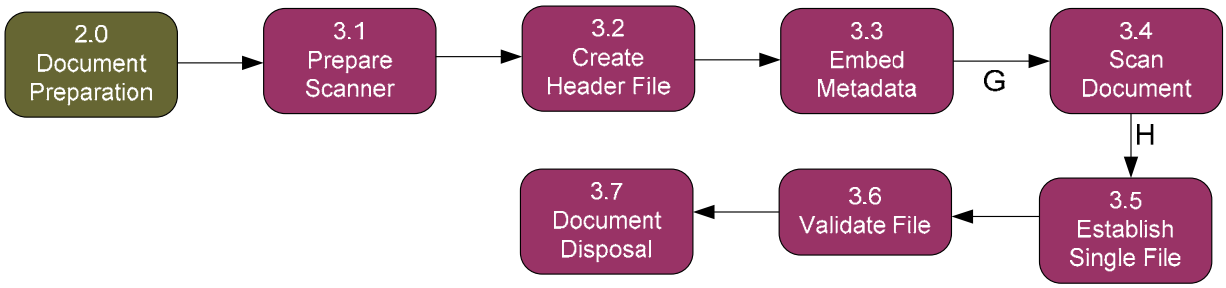

Figure 4-4. Digitize document sub-process

4.3.1 Prepare Scanner

The purpose of this step is to ensure the scanner and additional equipment being used for the scan is installed and working properly. Scanners should be calibrated using sample documents or images to measure elements such as tone, resolution, range, noise, and any additional characteristics that may be found to be important given a certain application (Billington, 2006). The information received from this helps the user determine if the proper scanner has been chosen for the situation as well as assisting the operator to ensure that the configuration and equipment controls have been set properly (Billington, 2006). The sample document should be selected based on situation and, and completed before work commences.

4.3.2 Create Header File

A header file for each document is now created. The header file is the cornerstone that makes step 3.5 of this sub-process possible, because the header file holds the links to all of the single page documents. This header file is later renamed to reflect the full document it represents. Each file should have specific information that defines it, such as the folder that contains it or the filing cabinet where has been stored. A majority of this information is obtained from the hierarchical structure created in step 1.3. This data about each file is identified and recorded so that it can be embedded into the header file in the next step.

4.3.3 Embed Metadata

Once the header file has been created, metadata needs to be embedded. Embedding the metadata is done using XML packets and storing them in the header file as a metadata stream (Adobe, 2001). Metadata streams are the preferred method of storage because they are more instantly recognized by search tools. To complete this process the XMP standards are followed and it is recommended that some kind of conversion tool is used to do the actual embedding.

A conversion tool allows a user to enter data about a particular document into a form and from there the system embeds that data into the header file. There are many different types of conversion tools and to ensure that performance meets needs, research should be done to decide which tool fulfills the needs of the organization.

4.3.3.G Embed Metadata with XMP

XMP is a set of standards created by Adobe that state the best information to store as metadata, as well as explain the best way to embed metadata into files. This is generally done by placing the information into a data information dictionary or metadata stream. According to the standards, the data information dictionary is attached to the end of the file, but the stream is fused inside the file. The metadata package is created either with a conversion tool especially for that purpose or it can be manually coded (Adobe, 2001).

4.3.4 Scan Document

After the scanner has been prepared the document is scanned. During the scanning process the paper document is placed on the scanner and scanned into the system to create a digital representation of the paper document. When each page of the document is scanned a pointer from the header file to the newly added file is attached linking the files with a Virtual Association. To achieve the proper results for this step the operator must check the quality of the scan in intervals to ensure that the calibration of the scanner has not become skewed (Billington, 2006) . Also, the operator must remember to check each page to confirm that the full document is being scanned and no pages are being skipped or folded causing an unreadable scan.

4.3.5 Establish Single File

After the scan has been completed, a single file is established. Instead of combining each individual file into a whole document, the best way to combine these individual files would be to link them all together. In this way the files are associated without having to actually merge them.

A Virtual Association is the combining of document with the use of pointers relating each file with its predecessor and successor. There are multiple benefits to using this method. The backing up of files is much simpler because instead of storing every page, only the file that has changed is saved. Additionally, amending files from other documents or locations can be achieved by simply adding a pointer from the header file to the requested file.

4.3.5.H Establish Single File with Virtual Association

Virtual Association is a conceptual technique to describe the way the individual files are linked. The header file stores links or pointers to all of the single files. When the header file is accessed, all links are opened and the single files are all combined dynamically into a viewer. This viewer would be a read-only, single run instance that is discarded when the header is closed.

4.3.6 Validate File

The final step of this process is to validate the scan. To validate the scan, the operator should view the file post-scan to ensure that all files of the document have been attached and are readable. In addition, a select test group should be taken to inspect that the metadata was embedded into the header files properly. If the system is working properly and no mistakes can be found, then the test group can be tested less often. If more documents are being found with errors, then the test group should be tested more often

If the conversion process is done correctly, documents are organized into a structured hierarchy and ready for users to access. The organization process from Section 3.0 is designed to correct a poorly organized system; thus, using the organization process after conversion ensures that the system does not become poorly organized. This step is the quality assurance step for the digitizing of the document confirming that all steps have been fulfilled properly, producing high quality digital documents. Once the documents have been validated, the header files are placed into the folder structure created in step 1.7.

4.3.7 Document Disposal

After the documents have been scanned into the system, there are two choices on what to do with these documents: destroy or replace them to their place of origin. This decision should be made based on whether the documents are archival or active. If the document is purely archival then it can be destroyed, but if the document is active then the original paper copy should be kept. The reason for this is that keeping an original copy is beneficial during system downtime so employees can still access the document.

4.4. DMS Requirements

4.4.1 Document Centralization

A centralized document repository should be created, allowing users to access the data from multiple locations. It provides a way for documents to be classified so that similar documents from different projects or offices can be located from a single source. The opposite is also true; it allows the finding of completely different document types associated to the same project or field of study. Furthermore, centralization of the documents provides a search method to locate documents that the user would have not known existed.

4.4.2 Selecting a Document Management System

Many different types of Document Management Systems are available on the market to manage files. Choosing the proper system can be complicated, and sometimes requires a consultant. Some critical issues to consider when choosing the proper system:

- Archival vs. Working Documents
- Organizational and Industrial Requirements
- System Costs
- Additional Required Costs

Archival vs. Working Documents

Documents that may not need to be changed at a later date, such as a final version of a report, are considered Archival Documents. These documents are saved for reference purposes and will normally not be needed again. A system that only deals with storing final documents is cheaper than systems that require storage and editing capabilities. If scanning a final document and storing it for later retrieval is your foremost concern, a simple and inexpensive DMS may be your best solution.

Working documents are documents that need to be revised on multiple occasions. Project materials such as: maps, reports, and many other text documents fit into this category. This type of system is needed when employees are editing the current version of a document. In this case a more advanced system is required so that the user can make changes and keep up with the changes made

Organizational and Industrial Requirements

A list of requirements, even if it is a small list, should be created. Employee requirements, internal technology requirements, industry standards, and customer requirements, if they use the system, should be included in this list. For example, if a system fulfills all of the needs set forth by your list of requirements but your technology department decides that it does not meet/follow the company's policies, the system conflicts with the policies.

Most industries have a set of rules that also need to be followed. A couple of examples include:

- The health industry must follow HIPPA compliance regulations.
- Also, all organizations are subject to requirements of the Sarbanes - Oxley Act of 2002.

Specific regulations pertaining to the industrial standards involved should be researched and compiled. Following these regulations is a crucial part of choosing a DMS.

System Cost

A system's price depends on the requirements of the organization and level of modification required. Simple systems that are ready to install out of the box have less functionality, but are sufficient for the needs of smaller companies. More complex systems grow exponentially more expensive. If employees from the system's manufacturer are required to come and install the system, the price rises significantly as a result. Newcomers to the DMS market are generally much less expensive than older established DMS companies, and frequently their products are

more cutting edge. Much time and money will be invested in converting paper documents and even current electronic documents into the system.

Research for choosing the correct system for an organization requires time and resources. It is time well spent, since implementing a system that does not serve the needs of a company can result in complete waste. If the needs of the company are simple, the decision can be made very quickly. If the needs are very complex, a consultant might be a good solution to find the best choice, or at least make several recommendations.

Additional Required Costs

Additional required costs that are incurred during implementation of a DMS include numerous system components, required hardware, acquiring a methodology, and developing or purchasing a controlled vocabulary.

Higher-end systems could require the hiring of someone to maintain the system. If the system is not being implemented for a large company, this probably would not be advantageous. If the company has IT staff, the system would be well within their skills to maintain.

Budgeting also needs to be done for all hardware and software that are needed to run and support the system. For example, higher end systems will have OCR functionality built in, while less expensive systems will not.

Licensing is another factor to take into consideration. Some systems require a fee per workstation, or employee. For the most expensive systems, this can reach prices above $5,000 per user, per year. Other systems charge the organization a flat licensing fee, which is generally much cheaper than paying per user.

4.5. Common Errors

In this section, common errors that occur during document conversion are explored. The errors are listed by symptom, and the probable cause of the symptom is then explained. Many problems concerning DMSs can be traced back to an oversight or mistake that occurred during design or implementation.

4.5.1 Poor Quality Documents

Glare spots

Glare spots on laminated documents are common. Process 2.2 eliminates the risk of glare reducing the readability of a document.

Obstructed Pages

Documents often times can be covered with post-it notes or other obstructions that decrease the converted document's readability. In this case, the most desirable solution would be to modify

the document with the additional information. If post-it notes must be used, place them so they do not cover any text in the document.

Skewed Document

The document is angled when placed inside the scanner. This occurs when the operator angles the document in the scanner feeder or on the scan bed. This error can occur in Sub-process 3.2. To prevent this error from occurring, the operator should ensure that the document is placed completely straight on the scanner or in the feeder.

Poor Quality Scan

One reason or cause is that the scanner is not properly inspected to fit the type of document being scanned, and the scan does not produce quality standard document. This step happens when the operator of the scanner does not complete step 2.4 defining the equipment or characterize the scanner to meet standards. This error can occur in process 3.1. In order to fix this error complete steps 2.4 and 3.1 in their entirety.

Incomplete Documents

In some cases incomplete documents are discovered in the system. This problem undoubtedly occurred in step 3.4. During scanning, a page was probably left out or skipped inadvertently. If possible, find the original hard copy of the document and examine it to find if the missing page is still with the document. If it is, re-scan the document.

4.5.2 Incomplete Search Results

Incomplete Thesaurus

Avoid inadequate development of a thesaurus. If a thesaurus is less complete than it should be, it results in narrower search results and requires more precise searching to retrieve the documents you want. The thesaurus can continue to improve once the system is in use by tracking what terms are used for searching, and then incorporating them into the thesaurus.

Improper Metadata

Overlooking important terms to be included in the vocabulary is a costly mistake that can occur early in the process. This produces incomplete search results and limited access to certain documents that have metadata embedded with these missing terms. If the metadata is entered correctly in step 2.5, the risk of this problem occurring should be greatly reduced.

4.5.3 Damaged Documents

Damaging a document during scanning should be avoided at all costs. The stability of an aging document should be evaluated before conversion. The precautions in process narrative 2.1 should greatly reduce this risk.

4.5.4 Incorrect Document Placement

Mistakes concerning the hierarchical ordering of terms might have occurred. This type of error results in the disorganization of documents when they are to be placed in the system. Avoid this by checking your hierarchical structure as it develops. Place the terms in a pyramid-shaped diagram as outlined in process 1.3 to visualize the relationships between the terms. This should make the process of determining whether or not terms are on the proper level of organization much easier. Risks are associated regarding the maintenance of the vocabulary as responsibility changes hands. Process 1.5 addresses how to fix this problem.

Another common mistake often made here is that the metadata was improperly defined. When the metadata was defined, not all of the areas of the hierarchy were covered adequately causing some of the documents not to be indexed properly. This error can occur in 3.3 and 3.4. To fix this problem, review Sub-process 3.3 along with the hierarchy to complete the list.

Section 5.0: Access Methods

A Document Management System does not fulfill its purpose unless the documents it contains are readily available. The purpose of the Access Methods section of this paper is to detail the best practices employed by users to retrieve documents stored in a Document Management System. This section discusses traditional methods of document retrieval such as searching and browsing capabilities, as well as newer tools that allow users to visually search for documents and better remember files previously accessed. The last topic discussed in this section are ways to combine these access techniques in order to create the best possible way to find documents needed.

5.1. Searching

Searching is an intuitive way for users to locate documents. The search engine can be broken down into three main areas. These are the search query, the document retrieval, and the document ranking. The language of the search engine, the layout and complexity of the search interface, and the thesaurus are all factors when determining how users structure their query. The search engine employs indexing as the method to retrieve documents requested by the user's query. The ranking of these documents is achieved utilizing algorithms designed for that purpose.

A generally used model in traditional retrieval systems details the three-way tradeoff associated with search engine performance. This tradeoff is between the speed of document retrieval, the recall ratio, and precision. The recall ratio is the number of relevant documents retrieved compared to the total number of relevant documents available. Precision describes the ratio of the number of relevant documents retrieved to the total number of documents retrieved. The general rule is that one of the options must be sacrificed in order to improve the others. This tradeoff triangle is even more apparent when the number of documents and users of a database increase (Takeda, 2000).

Balancing these tradeoffs when designing a search engine are important to the success of the search. When doing this, an organization should take their business goals and user preferences into consideration. For example, the typical web user wants his/her results fast, and he/she want to find the information he/she needs on the first page. Thus, when designing web search engines, speed and precision are the dominant evaluation criteria (Takeda, 2000).

5.1.1 Search Language

It is important to educate users about how the search engine works. Users who know the 'language' of the search engine can construct better search queries and therefore more efficiently

find documents they are looking for. It is good to include a help reference close to the search box detailing how users may more effectively utilize the search engine.

Most engines utilize Boolean logic. This means that the engine uses keywords to search through documents, as well as operators to define the relationship between the keywords. Examples of operators include AND, NOT, and quotation marks. The AND operator denotes that multiple terms must appear on a page. Having the NOT operator in front of a word means that that word must not appear on the page. Whereas a phrase is normally used to determine the most relevant terms to the query, a search phrase enclosed in quotation marks must be found in its entirety within the document. Operators can be used in conjunction to effectively enhance a search.

5.1.2 Simple Search

Most search engines have a simple interface including a single textbox for query entry and a search button for query submission. Keywords or phrases are entered into the search box in natural language, and once the query is submitted, the search engine retrieves documents related to the user's search. These documents are ranked according to the relevance of the document to the search query. The results generally are displayed in descending order of relevance, with an article title and abstract for the user to review. The user may then browse the results and select documents needed.

5.1.3 Advanced Search

An advanced search can be the fastest way to locate well-known documents. The advanced search uses search constraints to find documents with specific details. One such constraint is the scope of the search. Users can limit the search to only the titles or abstracts of documents within the repository rather than using a full text search. The advanced search can also allow users to specify metadata fields or categories of the document they are searching for. Examples include searching for the specific author, date created, known keyword or category.

The advanced search can also be useful for users not familiar with how the search engine works. This is because in general, the advanced search breaks down the search into multiple boxes. The user would only have one box in the simple search, and would have to be familiar with the engines operators to define a specific search. Examples of specific search boxes include a box for words that must be found, words that may be found, and words that must not be found in the documents retrieved.

5.1.4 Conventional Thesaurus

The conventional thesaurus contains all terms specific to an organization and the organization's documents. The conventional thesaurus is similar to the lexicon developed in C.3.6. The conventional thesaurus displays terms in a hierarchical fashion (Fowler, 1991). For example, if the documents all pertain to civil-engineering in the conventional thesaurus, one level of the

hierarchy might include the sub-disciplines. The next level of the hierarchy would be different aspects of each sub-discipline. In civil-engineering, the structure of the conventional thesaurus might appear as follows:

- Civil-engineering
    - Structural engineering
    - Geotechnical engineering
    - Traffic engineering
        - Trip Distribution
        - Trip Generation
        - Route Assignment
    - Hydraulic engineering
    - Material Sciences

The conventional thesaurus provides a tool for users to reference when defining queries and also can be used with the associative thesaurus described later. The hierarchical structure of the thesaurus helps users to visualize the associations between different terms before defining a concept map with an associative thesaurus.

The conventional thesaurus also provides users direct access to documents pertaining to a specific subject matter (Fowler, 1991). Refer to figure 5-1. For example, if a user was to select trip generation from the conventional thesaurus, the results would be any document dealing with trip generation contained within the traffic sub-discipline of civil-engineering. Use of the conventional thesaurus in this manner allows users to find documents related to a specific area of an organization and eliminate irrelevant results produced by searches using natural language.

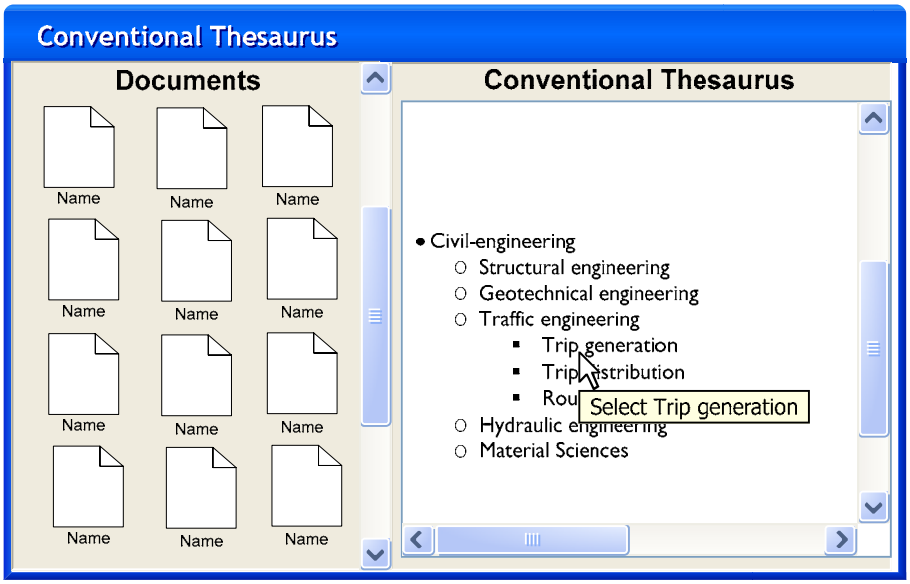

Figure 5-1. Accessing document through the conventional thesaurus

5.1.5 Graphical Queries

Queries allow users to input words in natural language. The query may consist of several words or entire sentences. The text entered in the query is then analyzed to identify words specific to the domain of the organization. Text analysis relies on statistical measures. The analysis must recover conceptual information from the natural language query. This is done by considering frequency and co-occurrence of words (Fowler, 1991). The identified words are used for the query.

The results of the query are displayed graphically using a conceptual map. The conceptual map consists of the nodes, which represent all documents that relate to one of the key words identified in text analysis. The more documents found associated with the word, the larger the node appears in the conceptual map. In addition to the size determined by relevance, a percentage of relevance appears by the node. The map also illustrates the relationships or associations between the nodes. Refer to Figure 5-2.

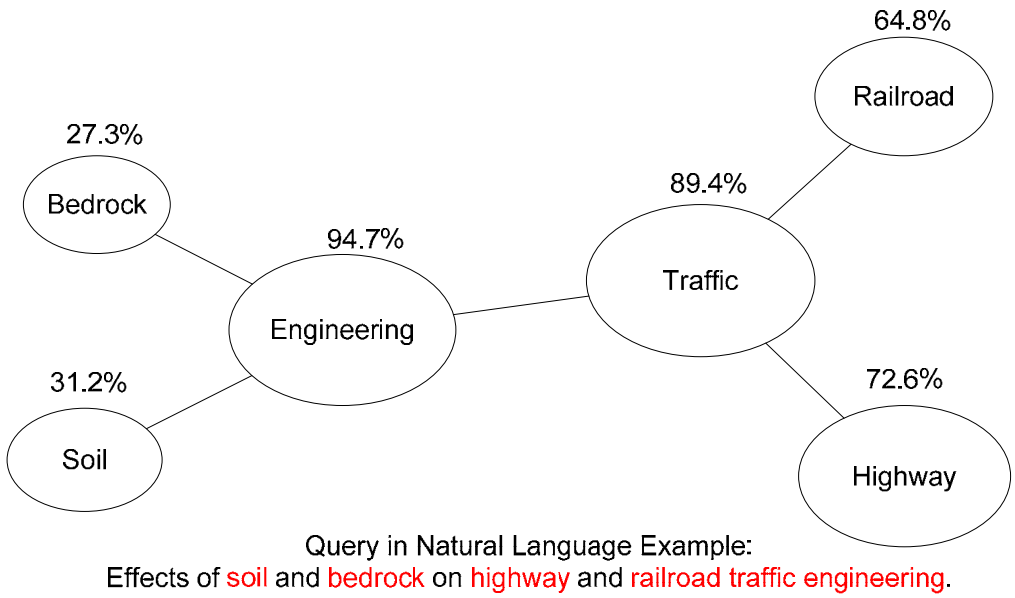

Figure 5-2. Query concept map

Users can refine their query graphically by selecting and deleting nodes in the map (Fowler, 1991). Selecting a specific node in a map allows the user to see a conceptual map of the documents within the node. Deleting nodes simply removes the node from the conceptual map and all documents contained within the node are not visible in the query results. The user is also able to refine the query by using the conventional thesaurus. The user can drag and drop words from the conventional thesaurus in order to further refine or broaden a query (Fowler, 1991).

5.1.6 Indexing and Retrieval

Indexing, when defined in terms of searching and retrieval, creates a data structure which allows the search engine to obtain results faster and more accurately. Essentially, a list of terms is created, wherein each term has pointers that detail where information about a document can be found. This list is compared to similar terms within the actual documents. The semantics of the words identify a document's main themes (Takeda, 2000). It is recommended to utilize document metadata as the means of indexing.

Search engines use the index to determine which documents to look into and where in those documents to look. Thus, the search engine does not search through the full text of every document to retrieve the results. In fact, the documents can be organized, stored, and indexed in such a way that promotes faster and more accurate retrieval (Lestina, 1997).

5.1.7 Ranking Retrieval Results

After a list of documents to be retrieved has been created, the order or rank of the documents based upon relevance to the search query must be determined. This is done using algorithms similar to the ones used to classify the documents for hierarchical purposes and metadata embedding. The major search engines generally use term weighting or vector space models, or variations of these algorithms.

Vector space models are similar to self-organizing maps discussed in the Technique Index. A vector is created for each document, and each coordinate of that vector is associated to a term or attribute of the document. The semantics of each term should distinguish that document from other documents within the search results, and terms or attributes that do not should be eliminated from the vector (Takeda, 2000).

A vector that represents the query is created in a similar manner. Documents are ranked by comparing their vectors to the vector of the search query. In simple terms, the ranking of a document is determined based on computation of the angular difference between the document vector and the query vector (Takeda, 2000).

The term weighting algorithm takes into account the how frequently a term appears within a document and the location of occurrence within the document (Takeda, 2000). The frequency is a large determinate of each term's ability to identify documents related to the query. Assigning weights provides a method of ranking documents. If a heavily weighted word indicates an association between the document and the query, an association is likely to exist; however, if a lightly weighted word indicates a relationship, the association between them is probably not as strong. Modifications and combinations of algorithms can provide better ranking of results.

5.1.8 Searching by Bit Configuration

Another way to find related files is to compare how they are stored in memory. If files have similar size and bit configurations, it is likely that they are close versions of the same file, or at least documents of a similar type. The UNIX operating system contains commands that allow the comparison of memory locations. Thus it would be feasible to create a search which compared two documents' bit configurations utilizing the header file's pointer locations.

5.2. Browsing

The purpose of browsing is to allow users to explore the hierarchy and locate documents through navigation. Retrieval systems should allow multiple mechanisms for document browsing to accommodate for different types of users. Browsing can be beneficial to both users who are familiar with the organizational structure and hierarchy and those who are not. Users familiar with the organizational structure can locate documents through direct browsing of the folder structure. Unfamiliar users can use the associative thesaurus to navigate and browse to locate documents.

5.2.1 Direct Folder Browsing

The purpose of direct folder browsing is to allow users that are familiar with the organizational structure and the naming conventions to locate specific documents quickly. The folder structure is directly related to the hierarchy. The structure is very organized with a standardized naming convention; thus, expert users can efficiently find documents through direct browsing.

Direct browsing is a more traditional technique of browsing the hierarchy and folder structure. Direct browsing is similar to browsing using Microsoft Windows Explorer. Direct browsing uses a two pane window divided vertically. The vertical window on the left displays the folder structure and is generally narrower than the window on the right. Each folder is selected and subfolders are displayed below in a hierarchical fashion. The right window displays all the documents contained within a folder. The user can select any of the documents displayed in the right window for viewing. Refer to figure 5-3.

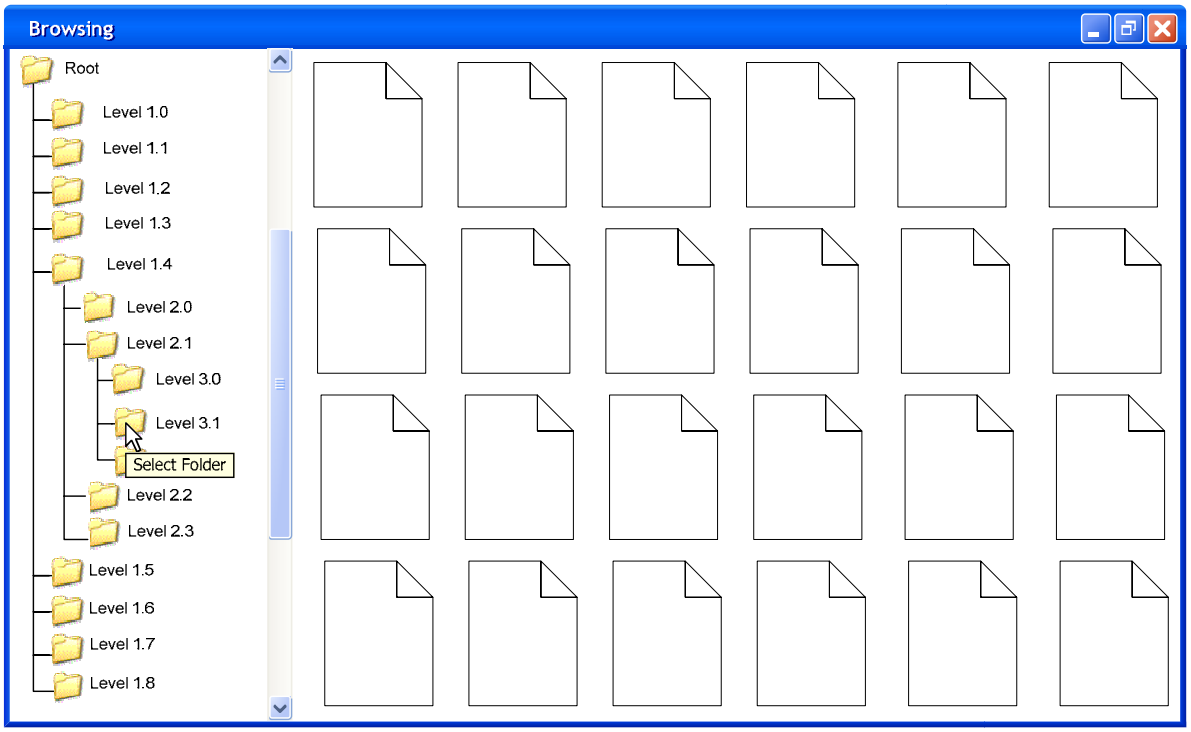

Figure 5-3. Direct folder browsing

5.2.2 Associative Thesaurus

The associative thesaurus helps users that are not familiar with the domain and may not know exactly what information they need. Similar to the conceptual map produced by a graphical query, the associative thesaurus gives a map of the entire organization. In the civil-engineering example from 5.1.5, the highest level of the associative thesaurus would be a node for each sub-

disciplines of civil-engineering. A user can select a sub-discipline and see a map of the organizational structure of the sub-discipline and documents that pertain only to that department.

As a user navigates further down in the associative thesaurus, the number of documents appearing in the document window is fewer because the user is reaching lower levels of the hierarchy. At these lower hierarchical levels, the documents are more specific with regard to subject matter. An associative thesaurus is a method of browsing that encourages users to explore the system (Fowler, 1991).

A user can also manipulate the associative thesaurus. The user can drag and drop terms from the conventional thesaurus to the associative thesaurus to see the associations between terms (Fowler, 1991). The dragging and dropping of terms in the associative thesaurus allows users to find very specific documents. Refer to figure 5-4 and 5-5. This method is similar to graphical querying; however, it is completely visual and does not require natural language. Thus, a user who is not familiar with the domain of the organization does not have to define the terms for the query.

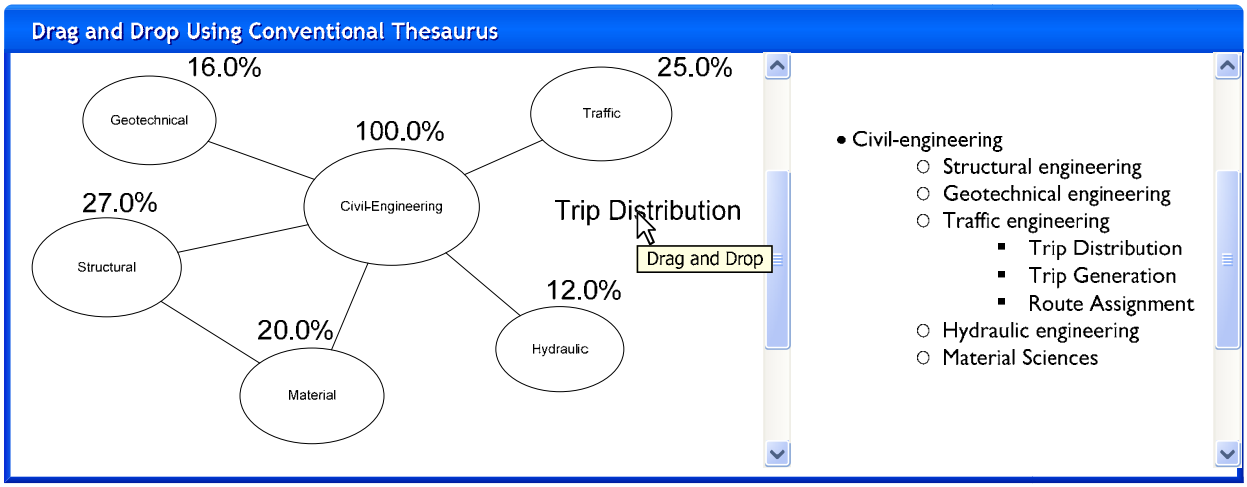

Figure 5-4. Dragging and dropping from conventional thesaurus

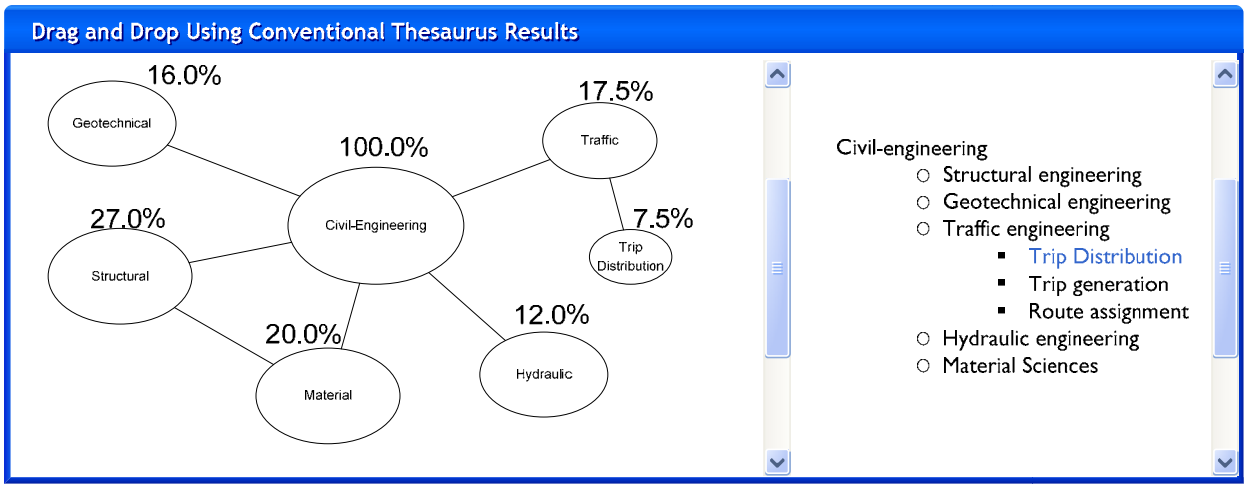

Figure 5-5. Dragging and dropping from conventional thesaurus

Notice in figure 5-4 that in the conventional thesaurus on the right-hand side of the screen the term Trip Distribution is highlighted in blue, and the user is dragging the term to the associative thesaurus window on the left hand side. In figure 5-5, the conceptual map has changed. The traffic node has become smaller because the documents pertaining to trip distribution were originally grouped in the traffic node but are now contained in the trip distribution node. Also, an association between trip distribution and traffic is visible. Because trip distribution is under traffic in the hierarchy, its association is to traffic. Just as in 5.1.5, the size of each node represents the number of documents associated with each node, while the percent relevance is displayed next to the node. In 5.1.5, relevance was referenced to the user-defined query; however, in the associative thesaurus, relevance pertains to the organizational hierarchy.

5.3. Visualizing Document Content

The purpose of visualizing a document's content is to allow more efficient use of documents and to help users identify relevant documents more easily. Visualization of document content allows users to scan large amounts of information relatively quickly (Fowler, 1991). In order to visualize documents more easily, a method of viewing document content without physically opening the file aids in faster navigation and identifying documents. Different methods can aid users in visualizing document content. Refer to figure 5-7. Thumbnails and abstracts assist users in visualizing document content and gaining bibliographical access to documents.

5.3.1 Thumbnails

Thumbnails are small images of the documents that provide a visual representation. Thumbnails generally depict the first page of a document, which is often a title page or contains a large heading. Thumbnail images are too small to read the text contained within the document; however, they provide a good depiction of the document layout. Different document types have

unique layouts and formatting; thus, a user knowledgeable in the domain or subject matter can recognize the layout and formatting of specific document types. To be beneficial to a user, the user must know the type of document he is looking for and must be knowledgeable in the domain.

Since the thumbnails are small images of the first page of the document, a method for more clearly identifying a document's layout and formatting is necessary. Users may hover over documents in order to see a larger image of the documents. Refer to figure 5-6. If a large number of thumbnails are present in the document window, allowing the user to see one thumbnail in a larger image helps the user to focus attention on that particular document. This helps prevent the user from becoming distracted by the other thumbnails.

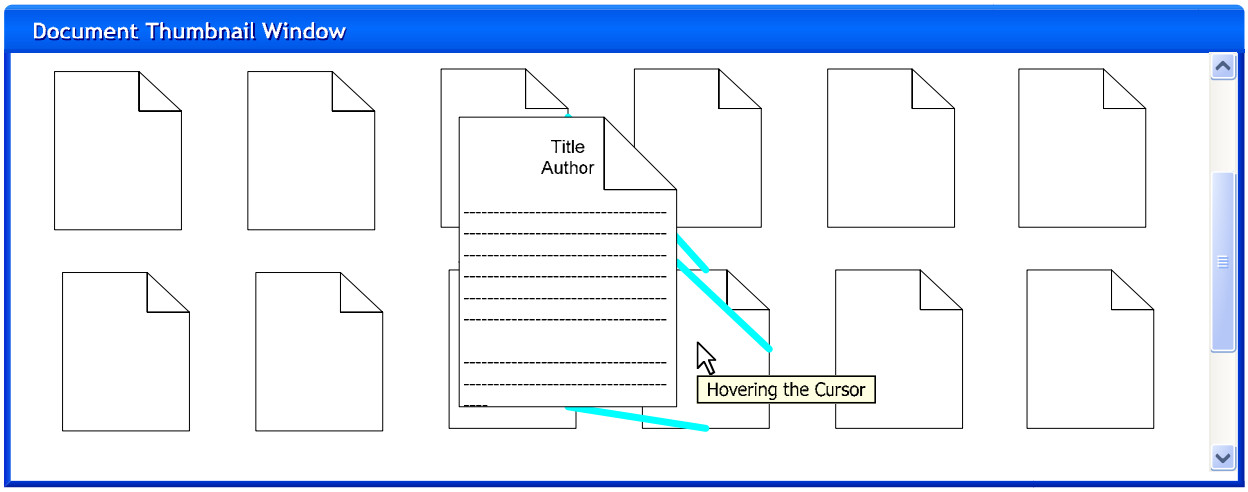

Figure 5-6. Enlarging document thumbnails

Searching and browsing can produce a large number of thumbnails in the results window. Many of the documents represented by the thumbnails may not be relevant to the user's information needs; thus, the user should have the option of eliminating the thumbnails that are not relevant. The user can select one or multiple thumbnails and choose to remove them from the results window. The user can also select any number of thumbnails that are relevant and only keep them in the results window.

This is accomplished through the use of check boxes. The user can check a number of thumbnails and choose to remove or keep the thumbnails. If a user chooses to keep the thumbnails, all thumbnails not checked are removed from the results window. Choosing to keep or remove thumbnails is not a permanent action. If the user runs the same search or browses to the same location later, all the thumbnails that were removed previously appear again. Allowing

the user to remove thumbnails keeps the user from becoming whelmed over by the number of thumbnails appearing in the results window.

5.3.2 Abstracts

The use of abstracts allows users to view some of a document's content without opening the document (Fowler, 1991). This is beneficial to both knowledgeable and unknowledgeable users. An abstract is a summary of text that describes the subject matter of a document. Providing an abstract of each document allows users to determine whether a document is pertinent to the subject matter they are looking for. When a user selects a document thumbnail, the abstract is provided in a small window. This allows the users to identify the document subject matter without opening the document.

The abstract contains the document's title and any relevant information about the author and versioning of the document. The title and versioning of the document can also allow users to identify if the document contains the subject matter they are interested in. The combination of the document summary and other relevant information regarding the document should be enough for users to determine the relevance of the document.

The abstract should also contain a list of all keywords associated to the document. This can be useful because it allows the user to search for keywords found in the document in order to find similar or related documents. In fact, the keywords themselves can be hyperlinks that, when clicked, navigate to search results for all documents containing that keyword. This is accomplished because the keywords identifying specific documents are based off of their embedded metadata.

Some documents do not have pre-written abstracts. Abstracts for these documents must be generated. A tool to automatically generate these abstracts could benefit the system. In order to achieve this, algorithms must be able to locate sentences whose subject matter define or explicitly state the purpose or overall 'message' of the document. These sentences must have weights assigned to them by the algorithm, and the most relevant passage or passages should be used to form the abstract. Finally, there must be minor adjustments to ensure that the abstract is cohesive; for example, words may need to be changed or removed to create a standard tense.

5.3.3 Bookmarking

The use of bookmarking allows users to save a link to a document they may need for later reference (Fowler, 1991). Once a user finds a document that is pertinent to the subject matter of interest, the user may bookmark the document. A bookmark does not physically save a document; rather, a pointer identifies the location of the document. The pointer allows the bookmarked documents to appear in a window as thumbnails, and the document abstract can be viewed by selecting the thumbnail. Using a pointer prevents the duplication of documents in multiple locations.

After a user has bookmarked a document, a new version of the document might become available; however, the pointer does not automatically update to the new version of the document. A user may prefer to continue working with the older version of the document, or may wish to begin using the new version; thus, the user must be alerted when a new version becomes available.

The use of Boolean (true or false) flags can accomplish this. When a document is updated, the new version has a true flag associated with it and the older versions flag is set to false. When a user opens the bookmark window, the pointers for each bookmark check the flag of the document to see if the document is the newest version. If the document is not the newest version, then the bookmark thumbnail is highlighted, which indicates to the user that it is not the newest version. The user may then view the newest version and decide which version to use. The user can then select to continue using the older version or use the newer version. If the user decides to use the newer version, the pointer is updated to point to the newer version.

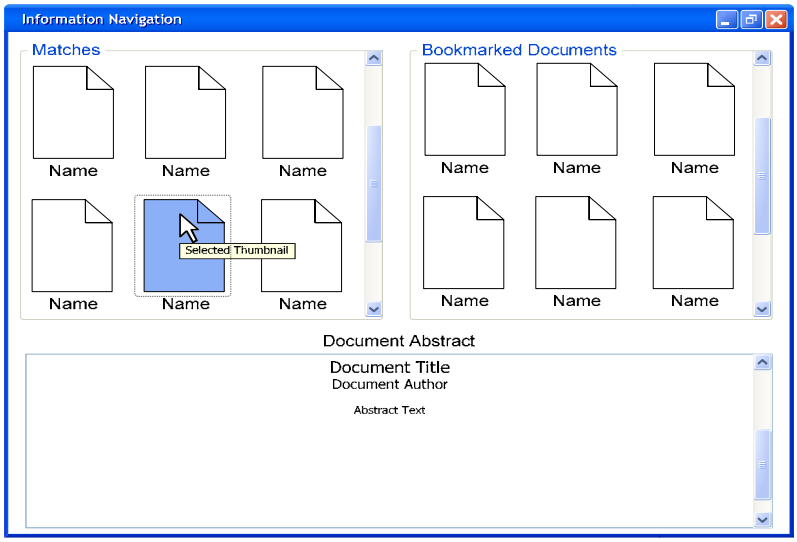

Figure 5-7. Viewing document content

5.4. EyePrint

EyePrint focuses on solving the problem of re-locating previously accessed documents. EyePrint attempts to solve three main problems associated with reusing digital documents. One is that digital documents do not contain clues that support information retrieval. Another is that it is hard to recall where a document was found. The third is that after an extended period of time, it is difficult to relocate documents (Ohno, 2004). For example, sometimes documents are moved and reorganized.

5.4.1 Traces

EyePrint uses traces which serve as visual cues to help people recall documents and recognize the documents they are looking for. A trace is simply a highlighted region or area on a document. The traces catch the attention of the viewer so the viewer does not need to read the whole document while trying to recognize it. The traces also become document attributes, which are used for searching (Ohno, 2004). An easy way to conceptualize a trace for digitized documents would be to consider a highlighted phrase on a paper document. The highlighted regions on a paper document draw the reader's attention. Refer to figure 5-8. Traces are designed to do the same thing for digitized documents. The traces also become embedded as metadata to aid in future searches (Ohno, 2004).

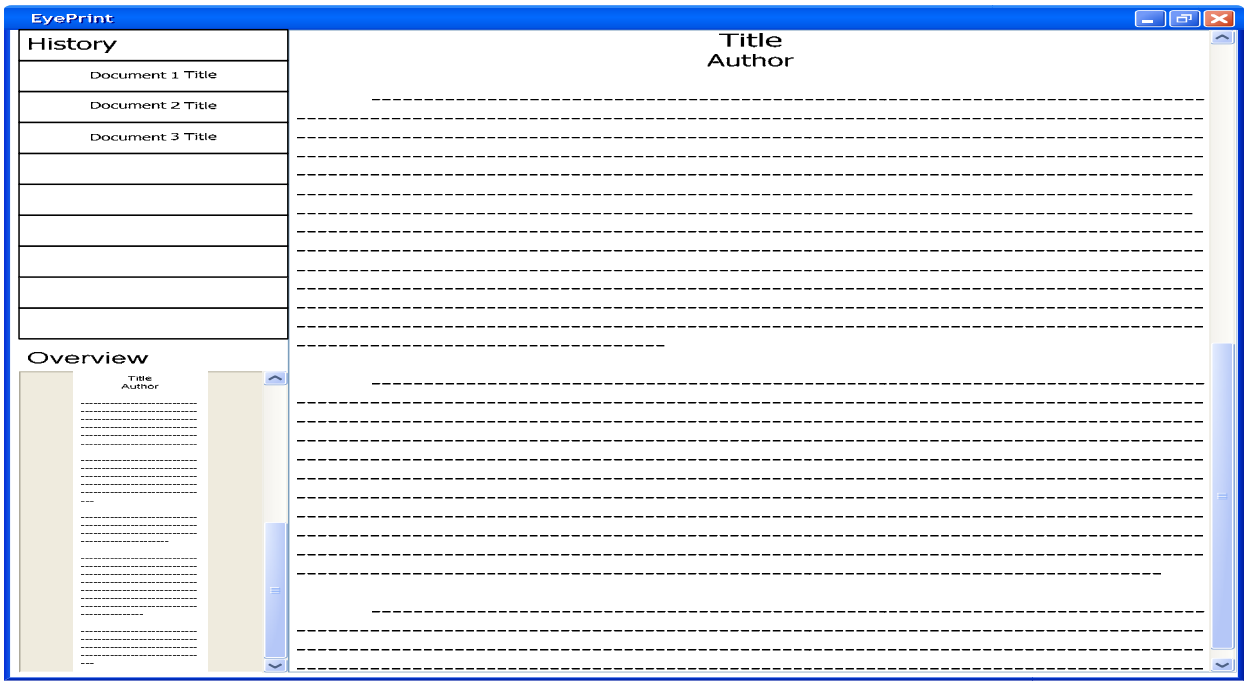

Figure 5-8. EyePrinting a document

5.4.2 Eye Gaze

The user's eye gaze generates the traces. A user's eye gaze indirectly represents their reading behavior. Users do not typically read the entire document and do not pay attention to what they have read; thus it is hard to remember what part of the document has been read before. The indication of the users reading habits or eye gaze in combination with the use of traces helps to identify the portions of the document that have not been read. Eye gaze concentrates on identifying the focus of the eyes on a portion of a document, while ignoring the portions of the document the user does not focus on (Ohno, 2004).

In order to identify eye gaze, the gaze tracking system is needed. The gaze tracking system uses very sensitive equipment to detect corneal reflection. The system takes into account the angle and calculates the position on the page the user is looking at. According to one study, the system is accurate within 0.8 degrees of the users view. However, the gaze tracking system does have downfalls. First, the users must keep their head in a fixed position or the system is not be able to measure the angle changes accurately. Second, the gaze tracking system is expensive (Ohno, 2004). Other measures of identifying eye gaze may be considered. For example, many users employ the cursor to follow along with their reading; thus, tracking the motion of the cursor on the screen could be an effective way of identifying traces. Another possibility is to let the users identify their own traces rather than using eye gaze.

5.5. Combining Access Methods

The purpose of combining different techniques of document access is to provide several ways to locate and access documents. Allowing the user multiple methods gives them the option of choosing the method that best fits their navigational needs (Fowler, 1991). For example, if a user is not entirely sure of the document needed, searching for the document might be the best navigational method. At the same time, whether the user is knowledgeable or not would determine the type of search to use. An unknowledgeable user would be better off using a graphical query. Combining different methods of access increases the efficiency of the users.

5.5.1 Graphical Queries, Associative Thesaurus, and Conventional Thesaurus

Users are able to use different methods of access together. For example, the conventional thesaurus can be used to help refine a graphical query while also aiding users to define queries in natural language (Fowler, 1991). The conventional thesaurus also aids users to identify documents in the associative thesaurus; however, while the conventional thesaurus is being used in conjunction with graphical queries and the associative thesaurus, graphical queries and the associative thesaurus cannot be used in conjunction with each other. The associative thesaurus is based on the hierarchical levels of the organization, while graphical queries identify documents that can span the entire organization; thus, the two different levels of access do not interact.

5.5.2 Thumbnails and Abstracts

Thumbnails and abstracts can be used in conjunction with any other form of access. Query results can be displayed as thumbnails and the abstracts available. When a user browses through documents, whether using direct folder browsing or the associative thesaurus, thumbnails and abstracts are available for the user. The thumbnails and abstracts also work with each other. When a thumbnail is selected, the abstract automatically appears in the abstract window and if the user chooses to move to the next abstract in the abstract window, the next thumbnail is automatically selected. Thumbnails and abstracts are great methods of visualizing document content and work well together and with other methods of access.

5.5.3 User Interfaces

The interface used can vary. The type of user interacting with the system largely relates to the interface. Different methods of access can be selected to display on the interface. For example, if a user is knowledgeable about the domain, the interface may only need to contain direct folder browsing, searching, thumbnail, abstract, and bookmark windows. Refer to figure 5-9. If the user is unknowledgeable about the domain, the interface might have windows for graphical queries, associative thesaurus, conventional thesaurus, thumbnails, abstracts, and bookmarks. Refer to figure 5.10. Users have the option of choosing the different access methods they wish to be displayed on their interface. Similar to viewing toolbars, the user can select different methods that are viewable and arrange the windows as they please. Allowing users this type of control over the interface is necessary because different users have different access preferences and these preferences may change over time.

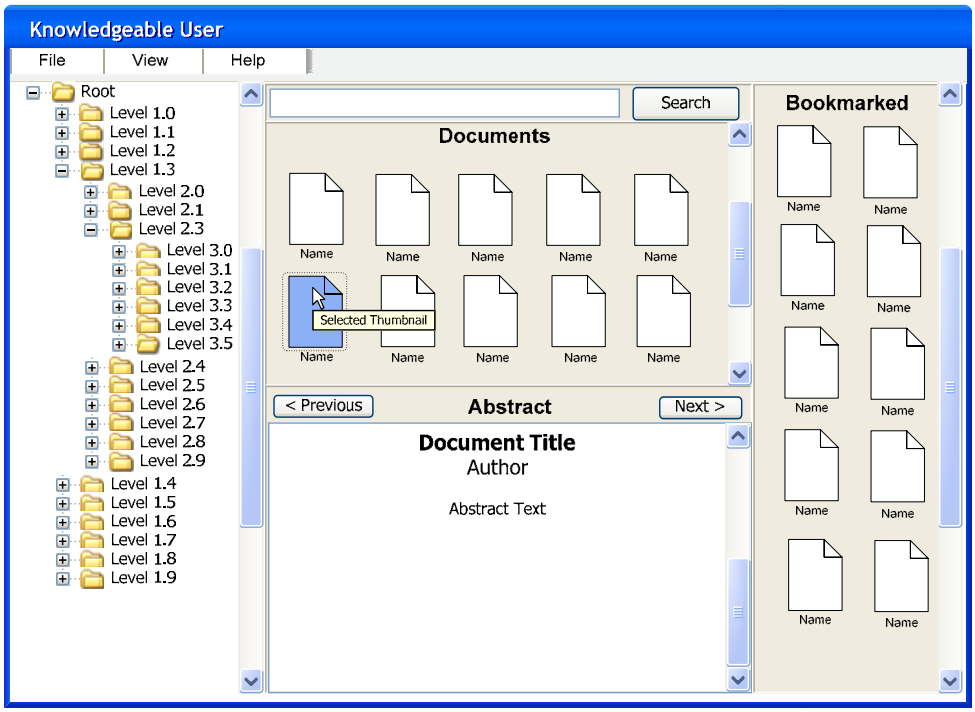

Figure 5-9. Interface for a knowledgeable user

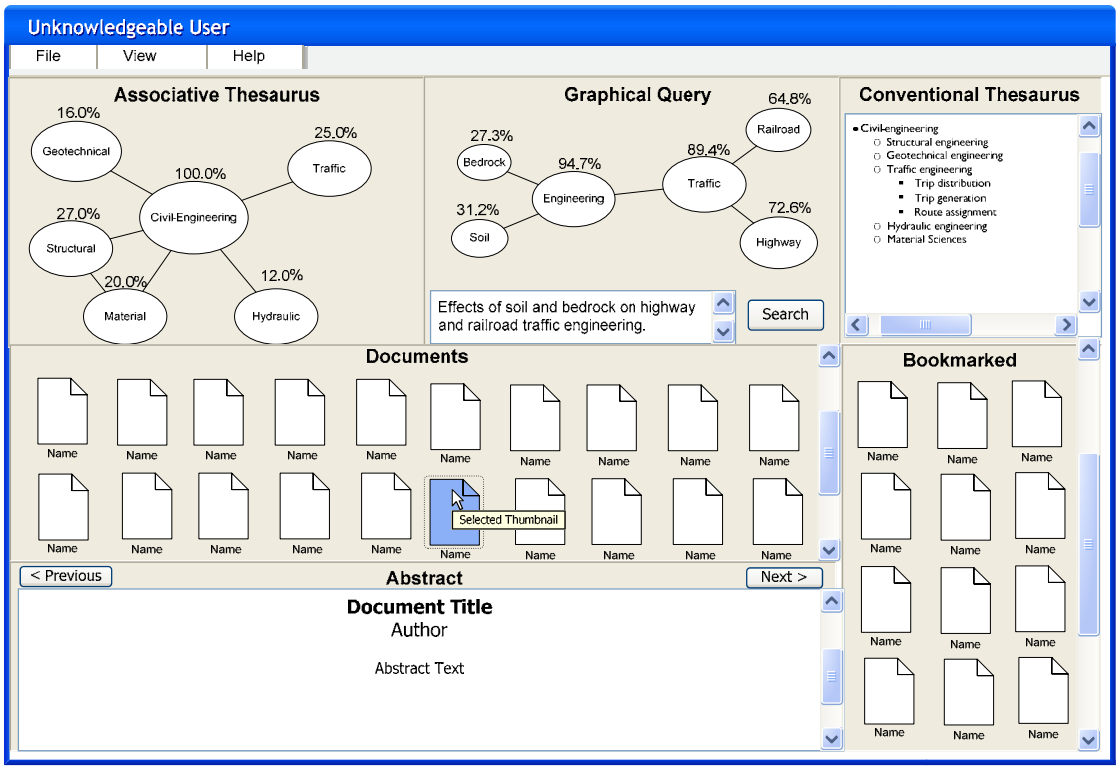

Figure 5-10. Interface for an unknowledgeable user

Section 6.0: Technique Index

The organization, conversion, and accessing process involved in Documents Management System (DMS) must employ different techniques and technology to successfully and efficiently manage an organization's documentation. Without the use of different techniques and technologies, good document management would not be possible. This section introduces different techniques and technologies for organizing, converting, and accessing documents in a DMS.

As with any process, cost is a major factor in determining the feasibility of the process. For each technique, potential cost sources are assessed on a relative scale in two areas. These two areas are Operational and Technological. The scale is made of five points, with a 5 being the most expensive and 1 being the least expensive relative to the techniques. It should be noted that those some of these techniques may be expensive, their cost is likely covered by the increased efficiencies each provides.

6.1 Optical Character Recognition (OCR)

The term OCR refers to a computer based software technology that uses complex algorithms to read characters off of an image of text. Images can be created in a variety of formats. Two of the typical formats for converting paper documents into electronic documents are TIFF and PDF, because they are generally higher quality that other formats. The general idea is to convert images of text into editable text, or in the case of this project, searchable text.

Current OCR engines are able to recognize typewritten text, hand printed numbers and alphabets, OCR generated characters, bar codes, magnetic ink characters (i.e. the routing number on a check), as well as formatting: spacing, checkboxes, etcetera. The downside is that current software does not recognize cursive text. If the characters are linked in any way, then they are rejected. OCR software bundles also often include several features. Noteworthy features with relevance to this project include the following:

- Batch processing.
- Despeckling/grayscale (which cleans up 'noisy" images or images with bad contrast).
- Template creation/field definition (allows the OCR to "know" where to find text).
- The ability to convert output to word processors, PDF, HTML, etc.

There are three methods for an OCR system to evaluate an image: full-page OCR, zone OCR, and dynamic OCR. Full-page OCR scans the entire image and pulls all recognized characters into the editable OCR file. Generally if the character is not recognized, then it is omitted,

although some OCR engines will keep the image of the unrecognized characters in the OCR file. Zone OCR restricts its search for data to specific, predetermined parts of the page. This can be useful when searching for headers or other specific document formatting, which in itself is valuable when searching for material that provides classification for indexing. Dynamic OCR (also known as Dynamic Forms Processing or Unstructured Forms Processing) searches for specific character sets anywhere on the page (Simple Software, 2007). Dynamic OCR can be in combination with full-page and zone OCR.

OCR engines can search for any recognizable characters, or use parameters to look for specific character sets or formats. Template and dictionary matching are methods used to accomplish this, and they can be used in combination with both dynamic and zone OCR. Template matching searches for any fields that match the template. For example, a social security number might follow the template ###-##-####, that is, a specific number of numerical characters separated by hyphens. When using dictionary matching, the OCR engine can be given a list of several terms or values to look for (Simple Software, 2007).

OCR engines are not perfect, and there are two error types generated. The first is rejection. When rejection occurs the system does not recognize the character at all. The characters that are rejected are included in the editable document as pictures; however, the user is unable to change them. The second error is that of substitution, or a "false positive". Substitution is an error produced when the system misidentifies a character and substitutes a different character in its place (Rosen).

Since errors do occur, ways to handle errors must be considered. Most all OCR software bundles allow an operator to review for errors, and even highlight words or characters with low identification confidence. That is, the software highlights probable errors. However, the substitution error can result in misidentified characters that still bear a high recognition confidence.

Industry acceptable error rates depend upon the material being processed by the OCR software. Going by today's standards, accuracy rates for typewritten text is above ninety nine percent. Acceptable rates for software reading handwriting range from eighty five to ninety five percent. However, a more advanced OCR engine (ACR defined below) purports conversion of ninety seven to ninety nine percent of handwritten material.

Another problem with OCR software is the learning curve associated with most products. What this means is basically, in order to recognize diverse handwriting/font sets, the system must first "learn" them. This often involves training sets which can be tens of thousands of characters long (Kahana, 2003). The bottom line is that in-house custom tailoring the OCR technology to fit a specific business need can be a difficult process that can end up taking several months.

Since molding OCR to a business need can take so long, it is important not to overlook the decision process regarding which OCR software package to use. As mentioned earlier, certain packages come with useful features, while others may not have features that are needed, which means that they have to be coded in with the system. It is our recommendation that a company seeking to use OCR should first contact several vendors to see if they will allow testing of engines. The engines should be compared for accuracy, the packages features should be compared, and costs of the packages should be compared as well. Once this comparison is completed, the company can better decide which OCR software package is best for their needs.

Other terms have been coined based off subsections of the OCR field. Since many typewritten text are not difficult for OCR software to recognize, the acronym ICR, or Intelligent Character Recognition, describes the more recent efforts to enable recognition of handwritten text. ACR is another recent term. Advanced Character Recognition, unlike ICR, does not require the handwritten text to be extremely neat and structured; rather, it can recognize "what is obviously diverse and difficult to read handwriting" (Kahana, 2003). Note that when this document references OCR, it includes the terms ICR and ACR if they apply.

Cost:

Operational: 2. The cost of utilizing OCR does not include the cost of scanning the documents as that is done with or without the using OCR. This makes OCR extremely low on the operational cost.

Technological: 3. OCR technology is relatively inexpensive. However, the use of ACR or ICR could increase the cost of this technique.

6.2. GIS

GIS is a geographic information system software with the capability of transforming paper map data into useable digital map coordinates. There are two common methods to use when digitizing maps into GIS. The first method uses a tablet and a digitizing puck to insert x and y coordinates into the system. This method is not considered the best because the tablet must be constantly re-calibrated to match up with map on the computer, and the user must watch the computer screen as well as the tablet to ensure they are in unison. The second, more preferred method is done entirely on the computer screen using materials already scanned into the system. This method is preferred because it results in a more accurate conversion, and it also allows the user to georeference the map using other pre-existing data.

Georeferencing is the process of converting a scanned image from pixels to x and y coordinates by either identifying places on the map with known x and y coordinates or by linking to corresponding locations in another georeferenced data layer. In other words, the map can be given coordinates by entering several specific locations' coordinates by hand, or marking several

locations and allowing the software to match those locations to the same places on a previously georeferenced document.

An example of the preferred method is provided in figure 6-1 below. The map from the article was georeferenced by stretching it over a previously georeferenced map of the same area. The points on the map are locations that correspond between the two maps.

Cost:

Operational: 4. Tracing each map into the system takes extensive time and knowledge of the maps. This makes this technique one of the most expensive to perform.

Technological: 4. The software and hardware necessary to perform this technique can be very expensive. Combined with the high operational cost, this is likely to produce the most cost for the least return of all the techniques. This is why it is considering to be a completely optional step and only advisable if mapping is a significant portion of the organization's operations.

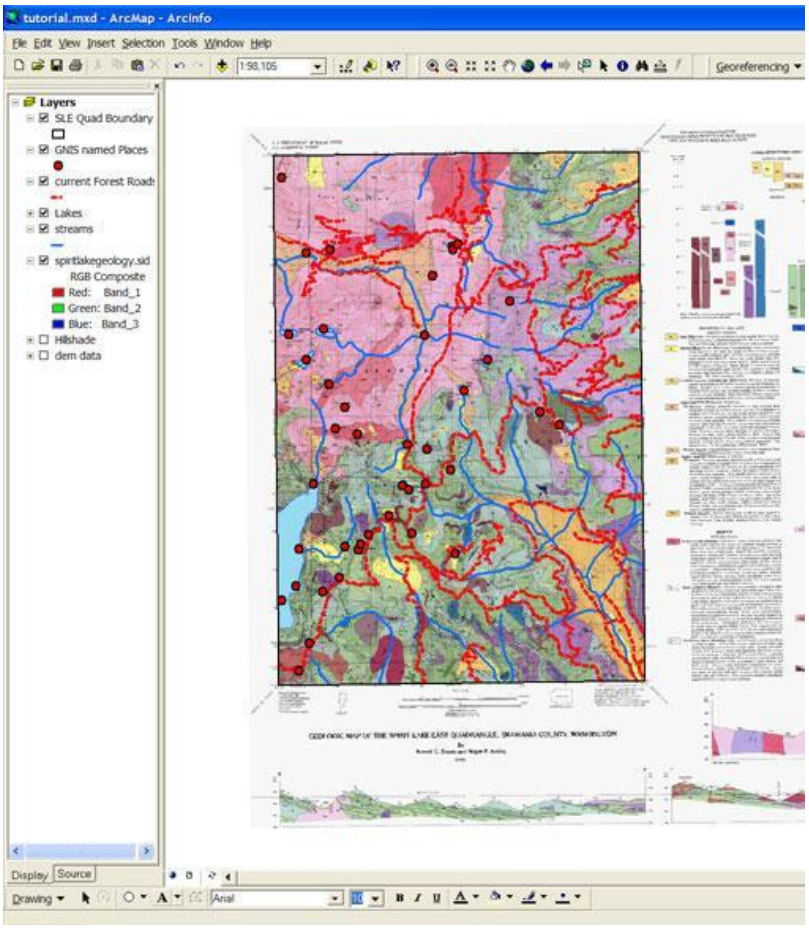

Figure 6-1. Georeferencing

6.3 Text Classification Algorithms

A text classification algorithm identifies the frequency words which occur within documents and associates a level of importance to each word. The level of importance is called a weight. The weight of each word is dependent on the frequency of its use. Frequency can affect the weight of a word in two ways (Caldas, 2002).

The algorithm identifies word frequencies with respect to the number of times a word occurs throughout one document. The more frequently a word occurs throughout one document, the more relative the word is to the classification of the document topic or classification. Words that allow users to discern the topic of a particular document bear a higher weight (Caldas, 2002).

The algorithm also determines the number of times a word occurs in all documents. The more a word occurs throughout all documents, the more poorly the word discriminates between documents (Caldas, 2002). Words that are frequently used in many documents bear a lower weight.

Both discriminatory words and words that cannot be used to classify a document need eliminating from the vocabulary to prevent unrelated documents from being associated with each other. Many different algorithms are available for text classification. Some of these algorithms include (Caldas, 2002):

- Support Vector Machines (SVM)
- Naïve Bayes
- K-nearest neighbor
- Rocchio's algorithm

Each algorithm produces its own percentage of accuracy due to different factors. A study showed SVM produced a 91.12% accuracy rating, which is comparable to human performance. The other algorithms performance ranged from 49-64% accuracy in the particular study (Caldas, 2002). Each classification algorithm has its own characteristics. For example, one variation of Naïve Bayes (one of the more popular machine learning methods) does not capitalize on term frequency weightings (Kim, 2006). Thus, when evaluating which classification algorithm to use, the characteristics of each should be analyzed. Documents have to be adapted to fit the specific representation and input formats of each algorithm; thus, the algorithm that already most closely fits current document representation should be chosen (Caldas, 2002).

Cost:

Operational: 3. Modifying each algorithm through repeated cycles to obtain the best possible performance takes an expert and requires much labor. However, once the algorithm reaches the desired goals, the amount of operational cost reduces significantly.

Technological: 3. Depending on the algorithm utilized, the technological cost varies. Each algorithm requires a different level of depth and complexity and may require an expert to develop.

6.4 Weighting Terms

The text classification algorithm identifies the terms used to classify documents; however, specified terms vary in degrees of importance. The identified terms must have weights assigned to indicate their importance. In the vector space model, the terms are stored in a vector or chain that represents all documents. The vector space model is a mathematical model used for indexing and relevancy rankings. In a vector space model, documents are represented as a vector of index terms or keywords. The vector is a list of the words not including the identified discriminatory words (Caldas, 2002). After the text classification algorithm has determined the frequency of the words, techniques have to be used to weight the term for relevance. Text classification algorithms and term weighting algorithms work hand in hand to classify documents. Different algorithms are available to weight terms. The following is a list of commonly used weighing algorithms (Caldas, 2002):

- Boolean weighting
- Absolute Frequency
- itc-weighting
- Entropy weighting
- tfxidf-weighting
- tfc-weighting

Each weighting technique produces difference result; thus, multiple methods should be evaluated to determine the best method in the particular situation is identified. The text classification algorithm plays a role in which weighting technique is the best. Once the terms have been weighted, they are referred to as index terms.

Cost:

Operational: 2. Weighing of the terms is a fairly automated process. The operational cost incurred is from checking the accuracy of the weights.

Technological: 3. Weighing terms plays hand in hand with the algorithm, making the cost of weighing terms vary with the cost of the algorithm, as each algorithm requires a different level of depth.

6.5 Lexicon Development

A lexicon (thesaurus) should be developed so the index terms are recorded and can be referenced. The lexicon not only should include index terms or key words, but synonyms and homonyms for the domain (Reul, 2005). The domain is the specific company, industry, field, etc. being studied. Recording the synonym allows matching of different words that mean relatively the same thing and provides for better clustering of documents. Homonyms should be taken into account in the lexicon because they are words that are spelled and sound the same but have different meanings (Reul, 2005). Homonym can affect the accuracy of document clustering; thus, homonyms should be avoided in the lexicon.

When developing the lexicon, each document must be parsed and the predefined words identified as index terms or key words are extracted. By predefining key terms, words insignificant to the process of indexing and clustering text are ignored. This helps prevent unrelated documents being associated with each other. This lexicon includes important parts of knowledge (Reul, 2005). As the process of general text analysis occurs, the lexicon should improve because of the frequency certain words are used. Frequency allows for the importance of certain words to be determined and for a structured hierarchy to be determined. The frequency refers not only to the occurrence of words in specific documents, but also the occurrence of words in all the documents which is determined by the text classification algorithm (Reul, 2005). The completed lexicon allows the development of a standardized hierarchy or product model.

The lexicon is different depending on the organization and industry to which the lexicon is pertains to. For example, when dealing with civil-engineering documents, you must have terms that are applicable to civil-engineering. Broad terms in a lexicon dealing with civil-engineering might be the sub disciplines, such as:

- Structural engineering
- Geotechnical engineering
- Traffic engineering
- Hydraulic engineering
- Material Sciences

The above terms would be the first stages of the lexicon. As stated above, as text analysis takes place, the lexicon should improve. To further illustrate the improvement of the lexicon in the example of civil-engineering, more specific words related to traffic engineering might be included in the lexicon. Such terms as:

- Trip Distribution
- Trip Generation
- Route Assignment

In the example of a civil-engineering lexicon, a basic hierarchy is beginning to form.

- Civil-engineering
    - Structural engineering
    - Geotechnical engineering
    - Traffic engineering
        - Trip Distribution
        - Trip Generation
        - Route Assignment
    - Hydraulic engineering
    - Material Sciences

Cost:

Operational: 5. Developing an appropriate lexicon takes an expert in the field, if not multiple experts. Their in-depth knowledge of the field is likely to come at a high cost. The high cost of their services, as well as this technique's heavy reliance on labor, make in the most expensive technique.

Technological: 1. Developing the lexicon requires minimal technological interaction and should be performed by an expert in the field.

6.6 Self-Organizing Maps and Topological Tree Structure

To achieve document organization, a self-organizing map (SOM) is used. A SOM is a vector or chain of documents linked together by their overall relationship to each other (Freeman, 2004). For example, an organization might have separate departments with their own documents; thus, each department is one part of the chain, however, all the departments are related because they belong to the same organization. The basic idea behind using SOMs is to categorize documents while maintaining the overall relationship between documents.

Grouping documents into topics based on discovered similarities (document clustering) helps achieve document organization. Detecting and tracking important topics over time also aids in the process of organizing documents. A self-organization map (SOM) in document organizations illustrates the advantages of using topological maps in identifying similarities between documents and clusters (Freeman, 2004). The use of the word "topological" refers to the properties of the documents that do not vary in the documents, or the collection of words that remain the same throughout associated documents.

The representation of a SOM allows for visualization of document clusters and the respective relationships between documents. Favoring important words in a cluster identifies or labels a cluster. Disfavoring words indicates the commonality of a word in all the documents or the

appearance of the word in unrelated clusters. Disfavoring words include the discriminatory terms and terms that indicate another cluster (Freeman, 2004).

Because the amount of documents in the entire collection is very large, a subset of documents is used to train the algorithm. The subset is called “training documents” and is a sample of documents that is representative of all the documents. Using a hierarchical model allows the general SOM to appear at the top and specific detailed SOMs at the bottom of the hierarchy structure. As each SOM becomes more specialized to a specific topic, navigation and visualization of the hierarchy structure is easier and more defined (Freeman, 2004).

The hierarchical SOM begins with a single map. Once the first map is defined, each node in the map is evaluated. The node consists of all the documents related to each other. The nodes are subsequently used to train child maps. A child map is a map made up of the documents in one node of the higher level map. The process of evaluating the nodes and breaking nodes into child maps continues until the lowest level is reached. The main advantage of this approach is that the number of terms is reduced in child maps, which become more specialized (Freeman, 2004).

The adaptive topological tree structure (ATTS) uses a set of hierarchically organized and independently spanned 1-dimensional growing SOMs (Freeman, 2004). Essentially, this means that at the highest level there is one vector consisting of all the documents. Each set of related documents make up a node in a vector or chain. Each node is broken down into any number of smaller vectors containing nodes representing a subset of related documents. Thus, each vector is independent of the other vectors at the same level of the hierarchy and is organized by relationships between documents (Freeman, 2004). Refer to Figure 6-2 for a depiction of a topological tree structure using SOMs.

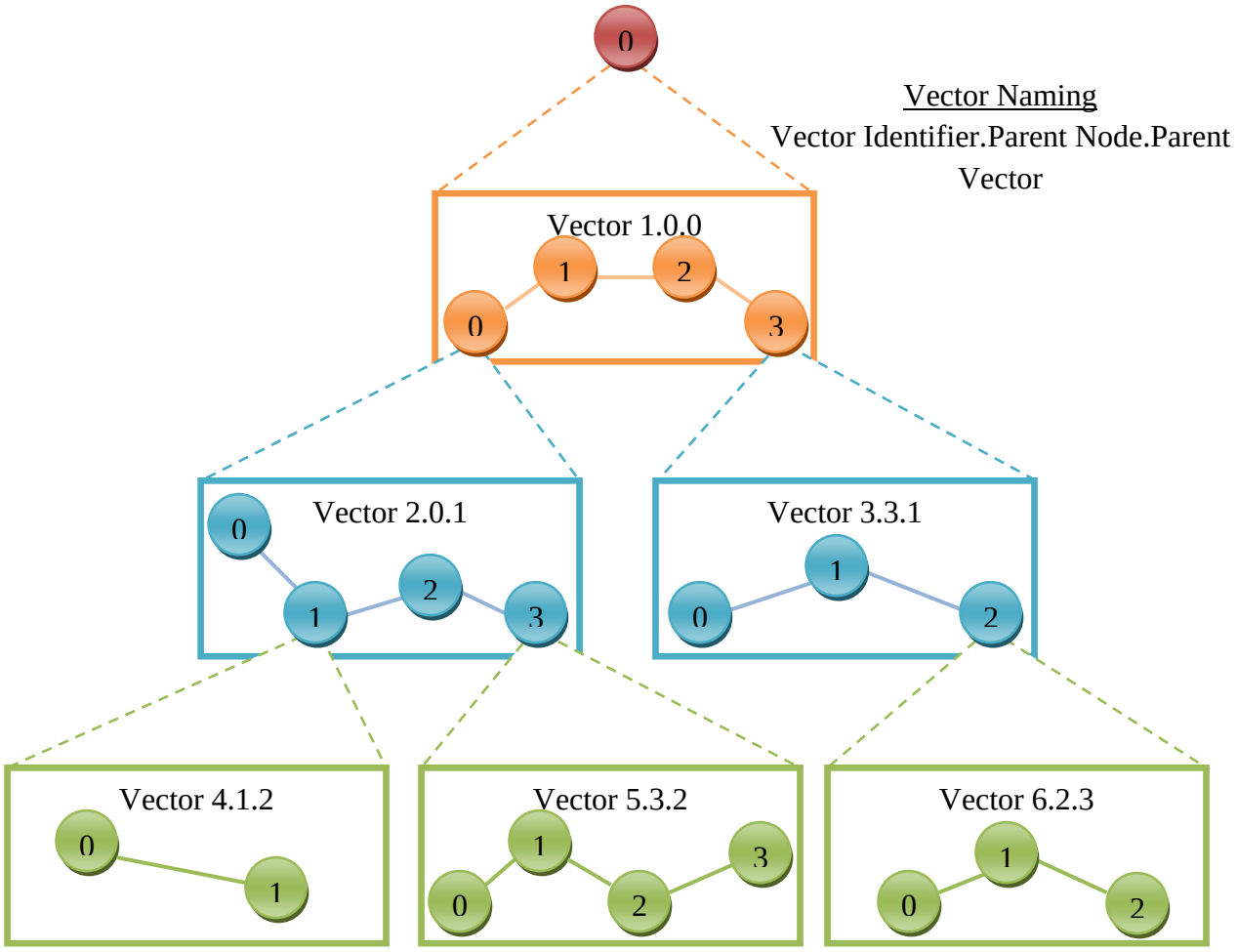

Figure 6-2. Tree structure with node expansion

The size of each vector or chain is independently determined through a validation process. The process of validation uses statistical measures to take into account the parameters (vocabulary) of the population and probability of distribution of the parameters. The statistical measures used rely on probability, frequency determined by text classification, and the weighting of each term (Freeman, 2004).

After determining size, documents clustered in a node are tested to determine if they are at the lowest level and cannot be broke down further, or if they should be further expanded into child chains. Because child chains have smaller vocabularies compared to their parents, the topic clusters are more specialized and specific. The addition of child chains forms a taxonomy. The taxonomy is arranged and organized according to the vocabulary defined for each level (Freeman, 2004).

There are different approaches to determining whether nodes should be broken down into further child chains. One approach to determine if a child chain is necessary is to allow a minimum number of documents in each child chain. Another test then checks the number of terms in the chain vocabulary. If the chain only contains few terms, this indicates that the parent node is sufficiently specialized, the documents may already be very similar, and there are insufficient terms for further classification. The final test is a method called term density test, which relates average frequency of terms to number of terms in the program dictionary (lexicon), to determine if further clustering is necessary. This ensures that there are a sufficient number of terms shared by the documents, and allows discrimination between them (Freeman, 2004). If any of these tests fail, then the parent node is at the lowest level, and no further break down is necessary.

There are several processes involved in the development of a topological tree, and the features of the tree offer benefits to end users. One such benefit comes from the associations developed between documents. These associations complement document names and help with exploration and navigation of the documents (Freeman, 2004). Refer to figure 6-3. After the initial parsing and indexing of the documents, developing the tree structure can begin.

First, terms must be selected and weighted using text classification algorithms and weighting techniques discussed in previous sections. Next, the documents must be separated into nodes according to their classification, which causes the SOM or vector to grow. The SOM must then be validated, and upon validation the hierarchy is expanded. After expansion of the hierarchy, if lower levels of the hierarchy are necessary, the process iterates back to selection and weighting. If no further levels are necessary, the vector or SOM is named, associations are defined, and the tree structure is complete (Freeman, 2004). Refer to figure 6-3.

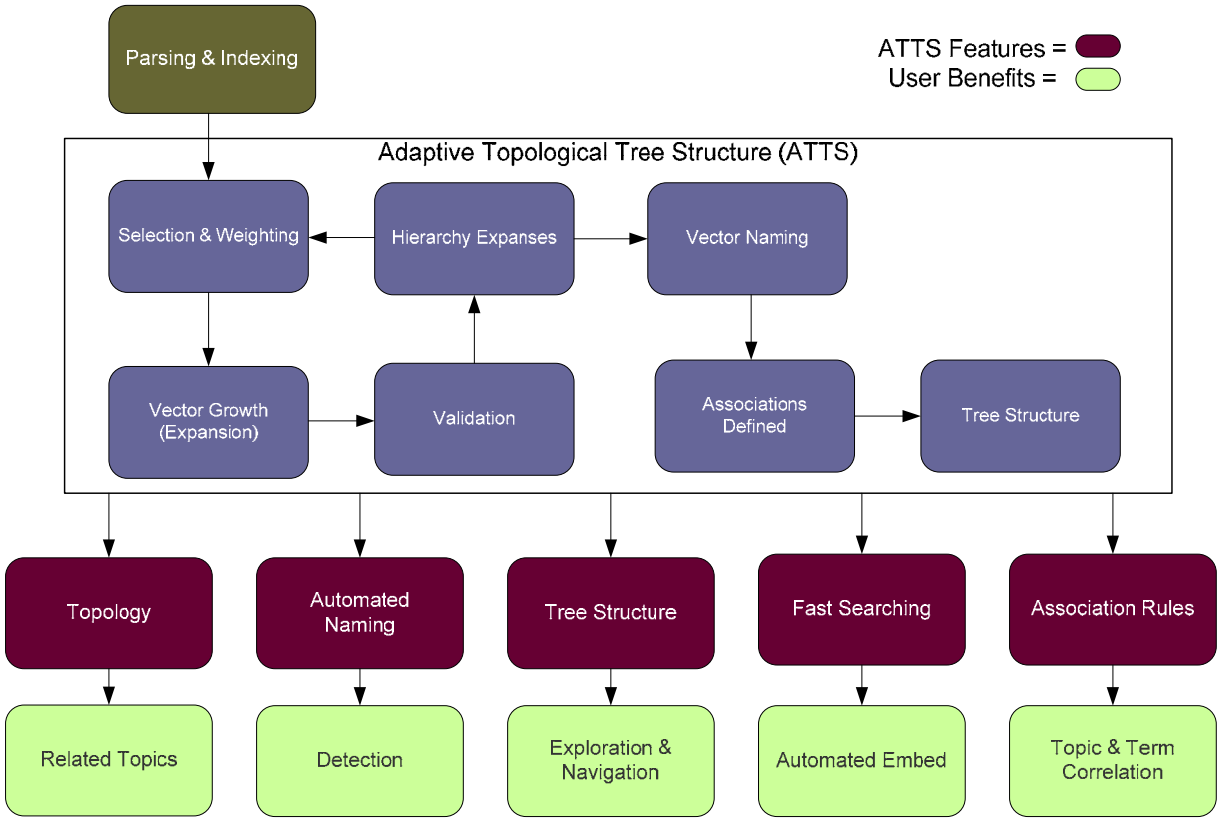

Figure 6-3. ATTS development, features, and benefits

Cost:

Operational: 1. This technique does not rely on human interaction, and if set up correctly could be fully automated.

Technological: 4. To be fully automated this technique would rely heavily on other techniques. Additionally, to make the tree structure would require sophisticated software.

6.7 Extensive Metadata Platform (XMP)

XMP, Extensible Metadata Platform, is a set of standards created by Adobe for the embedding of metadata into image documents. There are two main ways to incorporate metadata into PDF files. The first is called document information dictionary, and the second is known as the metadata stream.

The first method involves creating a document information dictionary. A document information dictionary, sometimes referred to as a metadata dictionary, is created by entering information about the document into an XML (Extensive Markup Language) packet and attaching it to the document. An XML packet is simply a way to implant XML code fragments into another file. This packet usually contains a list of information that identifies the document.

The metadata stream technique is similar to the document information dictionary, except the XML packet is fused within the document, as opposed to attaching it to the end of the file. Using a metadata stream is the preferred way to embed metadata into PDF documents for two main reasons. First, a metadata stream allows for the embedding of metadata-bearing artwork. In other words, images and other digital files that had metadata embedded by hardware during their creation still feature that accessible metadata. Secondly, the use of metadata streams gives search tools more effective and efficient methods of finding and determining classification of a document, thus PDF documents uploaded to the World Wide Web or within document management systems may be located, cataloged, or classified in less time.

After deciding which method is to be used, the actual attachment of the metadata stream or data dictionary must be performed. Two techniques for doing this are available. The first is to use a distiller program, and the second is to manually code the XML.

The distiller program prompts a user to enter information about the file in question into a form. Once all information is entered, the program embeds the data into the document as either a metadata stream or a data information dictionary, depending on which is selected.

Microsoft Notepad is an adequate application to use if using the manual coding approach. First, open the file in Microsoft Notepad and scroll to the bottom of the text information. Most of the file looks like unrecognizable garbage in Notepad, but near the bottom there is XML code that was created at the same time the file was first created. Enter the desired fields into this portion of the document to embed the metadata.

Cost:

Operational: 4. Determining the appropriate metadata to be associated with each file and then associating that metadata is very labor intensive. However, in the long run, having the correct metadata saves significant time and money.

Technological: 2. The cost of utilizing XMP is very low and doesn't require an expert.

6.8 Virtual Association

Virtual Association is a concept created to describe a specific way of linking electronic documents. It entails associating files without having to actually physically combine them. To do this, a header file is created that, when opened, dynamically opens and combines the contents of all files associated to it. Feasibly this could be accomplished using pointers to specific files. Most object oriented programming languages support header files and pointers to files. The data from the linked files would be combined into a temporary read-only viewer (possibly Adobe within a web browser) which would then be dismantled when the header file is closed.

It is important to note that just because a file is linked, does not mean the file is pulled from the pool of files. It is entirely possible that if a specific file has a high percentage match with two or more documents, that it may be included in both of those documents. This means that a single file may have multiple pointers that associate it to different documents.

The header file is simply a method of linking all the documents together. The header file does not hold any physical contents; it is a file that exists solely as a method of relating all files belonging to a document. The header file has several attributes. First, the header file points to all files that make up the whole document. Second, it has the order of the files. When the header file is opened it must merge the single page files, thus storing the order of the files provides a method of displaying the document in the correct order. The header file essentially represents the root of the files that belong to a document, while each file is a node. Refer to figure 6-4.

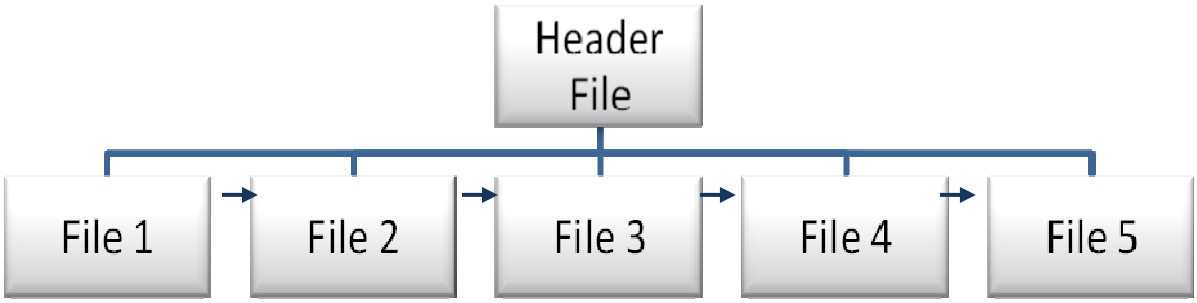

Figure 6-4. Header file pointing to all files and indicating file order.

As shown above, the header file indicates all files in a document and the sequential order of the files. There are distinct advantages to using a header file and virtual association. First, the use of header files makes error correction easier. If the files were physically combined into a whole document and one file did not belong in the document, the whole document would have to either be edited or deleted and recreated appropriately. The use of virtual association allows modification to only be made to the pointers within the header file and recreation of the document is not necessary.

Second, if users try to make alterations to the documents, the alterations are discarded after the header file is closed. The users are not actually viewing the files themselves; they are really seeing the representation of files as a whole document. Thus, any alterations made will not actually affect the files.

Cost:

Operational: 2. The cost incurred from determining which files need to be associated is fairly minimal as other techniques outlined in this section help with this process. Additionally, if the

conversion process is done correctly, the correct files will be scanned together, further reducing the operational cost.

Technological: 5. The virtual association techniques is new and needs to be further developed. Because of its experimental nature and it reliance on other techniques, it would be the most expensive technique.

Section 7.0: Conclusion

This project showed proper storage and effective retrieval strategies for document management systems. The project achieved this by integrating current best practices in the fields of civil engineering and information systems with leading edge practices from the document management community. A series of stages were implemented to confirm that the project achieved these objectives.

The first stage was to produce a literature review of current transportation specific document management systems. Section 1.0 of the project details document management systems in civil and transportation engineering. Another stage of the project was to study and benchmark departments of transportation all around the nation for best practices and unresolved issues. In Section 1.0 and 2.0 of the project benchmarked the DOTs experience with Data Management systems as well as the on-going document management activities at the Alabama Department of Transportation were studied to provide a proper framework for what the project should provide. Looking at the current state of the DOTs, it was shown that a tool was needed to properly capture what a digital document repository should accomplish. Sections 3.0 through 6.0 describe a process that, if properly implemented, provides an organization with a robust document management strategy that is effective, efficient, properly structured and reliable.

Overall the process described in the project takes a primitive document management organization and converts it to a more robust system which will make managing, storing and retrieving digitized documents a more effective and efficient part of the organization. Keeping with UTCA's theme of Management and Safety Transportation Systems, this project details a effective and efficient method of document retrieval. If an organization were to implement the process described here, decision making at the management level will be helped greatly by the proper timing and the reliability of the data this process provides. The process and findings of this report can be used to create a proper document management environment for any business, and more specifically transportation departments.

Section 8.0: Acknowledgements

This report was prepared with the cooperation and assistance of representatives of the following agencies and organizations: Alabama Department of Transportation; Federal Highway Administration-Alabama Division; Enterprise Integration Lab; Manufacturing Information Technology Research Center; Civil, Environmental, and Construction Engineering Department; and Area of MIS.

Section 9.0: References

Al-Deek, Haitham and Amr Abd-ElRahman. 2002. An evaluation plan for the conceptual design of Florida transportation data warehouse. Center for Advanced Transportation Systems Simulation (CATSS), 15-60-706.

Al-Deek, Haitham, Patrick Kerr, Balaji Ramachandran, Jeff Pooley, Adel Chehab, Emam Emam, Ravi Chandra, Yueliang Zuo, Karl Petty, and Ian Swinson. 2004. The central Florida data warehouse - phase (2). Center for Advanced Transportation Simulation Systems (CATSS), 16-50-713 New (16507011).

ATRC, Arizona Transportation Research Center. 2003. Evaluation of integrated document management system (idms) options for adot. Arizona Department of Transportation.

Back, W. Edward and Lansford C. Bell. 1995. Quantifying process benefits of electronic data management technologies. Journal of Construction Engineering and Management 121, no. 4: 415-421.

Balasubramanian, V. "Document Management and Web Technologies: Alice Marries the Mad Hatter." ACM Press Vol. 41, Issue 7(1998) 107-115. 02/16/2007 http://portal.acm.org/citation.cfm?id=278498&coll=>.

Billington, James. "Technical Standards for Digital Conversion." 12/21/2006. Library of Congress. 7 Mar 2007 <http://memory.loc.gov/ammem/about/techStandards122106.pdf

Bjork, Bo-Christer. 2001. Document management - a key it technology for the construction industry.

Bru¨ggemann, B. M., K.-P. Holz, and F. Molkenthin. 2000. Semantic documentation in engineering. In ICCCBE-VIII:828 - 835. Reston, VA: ASCE.

Caldas, Carlos H., Lucio Soibelman, and Jiawei Han. 2002. Automated classification of construction project documents. Journal of Computing in Civil Engineering 16, no. 4: 234 - 243.

Chau, K.W., Ying Cao, M. Anson, and Jianping Zhang. 2002. Application of data warehouse and decision support system in construction management. Automation in construction 12: 213 - 224.

Choi, K. and T. J. Kim. 1994. Integrating transportation planning models with gis: Issues and prospects. Planning, Education and Resources, no. 13: 199 - 207.

Cohn, Fred. 1995. New York gets wired. Civil Engineering 65, no. 9.

Corey, M. J., M. Abbey, I. Abramson, and B. Taub. 1998. Oracle 8 data warehousing: A practical guide to successful data warehouse analysis, build and roll-out.

Detwiler, Jim. "Digitizing." GIS Resource Document. 2002. Pennsylvania State University. 7 Mar 2007 <http://www.pop.psu.edu/gia-core/pdfs/gis_rd_02-39.pdf>.

"Embedding XMP Metadata in Application Files." Cover Pages. 01 September 2001. Adobe. 7 Mar 2007 <http://xml.coverpages.org/XMP-Embedding.pdf>.

Fast, Karl, Fred Leise, and Mike Steckel. "Creating a Controlled Vocabulary." 07 Apr 2003 <http://www.boxesandarrows.com/view/creating_a_controlled_vocabulary>.

Fowler, Richard. "Integrating query thesaurus, and documents through a common visual representation." ACM Press (1991) 142-151. 02/13/2007 <http://delivery.acm.org/10.1145/130000/122874/p142-fowler.pdf?key1=122874&key2=0332413711&coll=GUIDE&dl=GUIDE&CFID=11422057&CFTOKEN=34223574>.

Freeman, Richard. "Adaptive Topological Tree Structure for Document Oganisation and Visualisation." Neural Networks Vol. 17(2004) 1255-1271. 01/23/2007 <http://images.ee.umist.ac.uk/hujun/pubs/ATTS-NN-2004SI.pdf>.

Fruchter, R. 1999. A/e/c teamwork: A collaborative design and learning space. Journal of Computing in Civil Engineering 13, no. 4: 261 - 269.

Goodchild, Michael F. 2000. Gis and transportation: Status and challenges. GeoInformatica 4, no. 2: 127-139.

Guebert, A. A. 1991. Geographic information systems: Applications in municipal traffic engineering, ed. Transportation Association of Canada:C75 - C91. Ottawa, Canada.

Hajjar, Dany and Simaan M. AbouRizk. 2000. Integrating document management with project and company data. Journal of Computing in Civil Engineering 14, no. 1: 70 - 77.

Halkidi, Maria. "On Clustering Validation Techniques." Journal of Intelligent Information Systems Vol 17, No. 2(2001) 107-145. 01/11/2007

Kahana, Paz. "Advanced Character Recognition." ACR Technology and Performance Analysis White Pages (2003) 1-16. 01/26/2007 <http://www.charactell.com/ACRWhitePaper.pdf>.

Kim, Sang-Bum. "Some Effective Techniques for Naive Bayes Text Classification." IEEE Transactions on Knowledge and Data Engineering Vol. 18, No. 11(2006) 1457-1466. 01/22/2007 <http://web.ebscohost.com/ehost/detail?vid=4&hid=4&sid=450dea06-4d9c-4f76-aadb-37b5df305580%40sessionmgr3 >.

Kimmance, J. P., M. P. DBradshaw, and H. H. Seetoh. 1999. Geographical information system (gis) application to construction and geotechnical data management on mrt construction projects in singapore. Tunneling and Underground Space Technology 14, no. 4: 469 - 479.

Kosovac, B., T. Froese, and D. Vanier. 2000. Integrating heterogeneous data representations in model-based aec/fm systems. In Conference on Construction Information Technology, International Council for Research and Innovation in Building and Construction:556-566. Rotterdam, The Netherlands.

Kwon, Taek M., Nirish Dhruv, Siddharth A. Patwardhan, and Eil Kwon. 2003. Common data format archiving of large-scale intelligent transportation systems data for efficient storage, retrieval, and probability. Transportation Research Record 1836, no. 03-3519.

Latimer, DeWitt IV and Chris Hendrickson. 2002. Digital archival of construction project information.

Lauzon, R. G. and J. M. Sime. 1993. Connecticut's bridge management information system. In Characteristics of bridge management systems. Austin, Texas.

Lefchik, Thomas E. and Kirk Beach. 2006. Development of national geotechnical management system standards for transportation applications.

Lestina, Gregory. "Providing Document Retrieval Though Metadata Repository at the Census Bureau." U.S. Bureau of the Census (1997) 428-433. 02/16/2007 <http://www.amstat.org/sections/srms/Proceedings/papers/1997_072.pdf>.

Liu, Henry X., Rachel He, Yang Tao, and Bin Ran. 2002. A literature abd best practices scan: Its data management and archiving. Wisconsin Department of Transportation, 0092-02-11.

"More info on SimpleOCR." Simple OCR. Simple Software. 1/21/2007 <http://www.simpleocr.com/Info.asp>.

Munteanu, Dan. "A Survey of Text Clustering Techniques used for Web Mining." The Annal of Dunarea De Jos Unversity of Galati Fascicle III (2005) 54-59. 01/22/2007 <http://www.ann.ugal.ro/eeai/archives/2005/Lucrare-09-DanMunteanu.pdf>.

"Naming Conventions for Electronic Documents." Information Management. August 2005. Alberta Government. 7 March 2007 www.im.gov.ab.ca/publications/pdf/DocumentNamingConventions.pdf>.

Ohno, Takehiko. "EyePrint: Support of Document Browsing with Eye Gaze Trace." ACM Press (2004) 16-23. 01/22/2007 <http://delivery.acm.org/10.1145/1030000/1027937/p16-ohno.pdf?key1=1027937&key2=5595413711&coll=GUIDE&dl=GUIDE&CFID=11422057&CFTOKEN=34223574>.

O'Packi, Paul and Simon Lewis. 1998. What can a spatial data warehouse do for a transportation agency? In Geographic Information Systems for Transportation Symposium. Salt Lake City, Utah.

Paice, C. D.. "The Automatic Generation of Literature Abstracts: an Approach Based on the Identification of Self-indicating Phrases." (1980) 172-192. 03/05/2007 http://portal.acm.org/citation.cfm?id=636680&coll=portal&dl=ACM&CFID=12970945&CFTOKEN=76725238>.

Papiernik, Daniel K., Dhruv Nanda, Robert O. Cassada, and William H. Morris. 2000. Data warehouse strategy to enable performance analysis. Transportation Research Record 1719, no. 00-0603: 175 - 183.

Poe, V., P. Klauer, and S. Brobst. 1998. Building a data warehouse for decision support.

Proffit, Merrilee. "Pulling it all together: use of METS in RLG cultural materials service." Library Hi Tech Vol. 22, No. 1(2004) 65-68. 01/26/2007 <http://www.ingentaconnect.com/content/mcb/238/2004/00000022/00000001/art00007>.

Quiroga, Cesar A. and Darcy Bullock. 1996. Geographic database for traffic operations data. Journal of Transportation Engineering 122, no. 3.

Rasdorf, William, Kent Taylor, and Larry Wikoff. 2001. Data warehouse digital archiving case study: North Carolina department of transportation. Transportation Research Record 1769, no. 01-0119: 71-78.

Reul, Sabine. "Retrieval of Project Knowledge from Heterogeneous AEC Documents." Institute for Construction Informatics (Technische University Dresden) (2005) 1-7. 01/22/2007 <http://www.cib.bau.tu-dresden.de/projects/dokbau/lit/20000412_bauDoc_ICCCBE-VIII_dsr.pdf>.

Rosen, Richard. 2003. "Data Capture using FAX and Intelligent Character and Optical Character Recognition (ICR/OCR) in the Current Employment Statistics Survey (CES)." 7-14. 01/26/2007 <www.fcsm.gov/03papers/Rosen.pdf>.

Sanvido, V. E. and D. J. Medeiros. 1990. Applying computer integrated manufacturing concepts to construction. Journal of Construction Engineering 116, no. 2: 365 - 379.

Scherer, R. J. and S. Reul. 2000. Retrieval of project knowledge from heterogeneous aec documents. Computing in Civil and Building Engineering: 812 - 819.

She, T. H. 1997. Development of a geographic information system (gis)-based bridge management system (bms), University of Salford.

She, T. H. and G. Aouad. 1996. Development of an information model for a geographic information system - based bridge management system (bms). In International Conference on IT in Civil and Structural Engineering Design:103 - 111. University of Strathaclyde, Glasgow, Scotland.

She, T. H., G. Aouad, and M. Sarshar. 1999. A geographic inforamtion system (gis) - based bridge management system. Computer-Aided Civil and Infrastructure Engineering, no. 14: 417 - 427.

Simoff, S. J. and M. L. Maher. 1998. Ontology-based multimedia data mining for design information retrieval. Comuting in Civil Engineering: 212 - 223.

Smith, Brian L. and Simona Babiceanu. 2004. Investigation of extraction, transformation, and loading techniques for traffic data warehouses. Transportation Research Record 1879: 9 -16.

Stokes, R. W. and G. Marucci. 1995. Gis for transportation: Current practices, problems, and prospects. ITE 65, no. 3: 28 - 37.

Takeda, Koichi. " Information retrieval on the web ." ACM Computing Survey Archive Vol. 32, Issue 2(2000) 144-173. 02/19/2007 <http://portal.acm.org/citation.cfm ?id=358934&coll=portal&dl=ACM&CFID=16139484&CFTOKEN=98216933>.

Tate-Glass, Martha J., Rob Bostrum, and Greg Witt. 1999. Data, data, data - where's the data? no. A1D09.

Turk, Z., Bo-Christer Bjork, C. Johanson, and K. Sevenson. 1994. Document management systems as an essential step towards cic. In The CIB W78 workshop on computer integrated construction, ed. M. Hannus and K. Karstila, 8:22 - 24. Technical Research Center of Finland, Espoo, Finland.

Turner, Shawn M. 2001. Guidelines for developing its data archiving systems. Texas Transportation Institute, 2127-3.

Turner, Shawn M. 2002. ITS data archiving: Summary report. Texas Department of Transportation, 2127-S.

Wood, W. H. 2000. The development of modes in textual design data. Computing in Civil and Building Engineering: 882 - 889.

Xiong, Demin and Hui Lin. 2000. Spatial data handling for its: Perspective, issues and approaches. GeoInformatica 4, no. 2: 215 - 230.

Yang, M. C., W. H. Wood, and M. R. Cutkosky. 1998. Data mining for thesaurus generation in informal design information retrieval. Computing in Civil Engineering: 189 - 200.