Document Database Considerations

Richard J. Long, P.E. and
Andrew Avalon, P.E., PSP

LONG INTERNATIONAL

Long International, Inc. • 5265 Skytrail Drive • Littleton, Colorado 80123-1566 USA • Telephone: (303) 972-2443 • Fax: (303) 200-7180 • www.long-intl.com

Document Database Considerations

Table of Contents

Document Database Considerations

INTRODUCTION

Why Use Computerized Database Support?

A lack of knowledge on key facts or disputed issues may have significant adverse consequences in an arbitration or litigation. Computerized database support enhances the efficiency of document organization, research, and retrieval tasks performed by lawyers and experts by making these tasks easier, faster and more accurate. When searching for key information, for example, it is far more efficient to search a database and spend a couple of minutes locating key documents than it is to manually review potentially hundreds of pages to find those that fit the search criteria. If the document population comprises thousands of documents, a computerized database solution is essential to cost effectively search for relevant documents.

Benefits to Using Computerized Database Support

There are many benefits to using computerized database technology in conjunction with pending arbitration or litigation. Some of these benefits include:

- *Safety and Security of Documents.* Once a document is coded and imaged and an electronic copy of it is stored, it is virtually impossible to misplace or lose. A computerized database eliminates filing errors because a document may be located and retrieved using various search criteria within the database. Restricting the individuals who may gain access to the electronic document files enhances the safety and security of documents. Rather than housing documents in file cabinets or boxes in a room that may have minimal security, imaged documents are much more difficult to access because the user will need to clear system security and understand how to retrieve a document.

- *Reduced Storage and Photocopy Costs.* Images of documents are saved on discs that, depending on the type of media, may have the ability to store thousands of pieces of paper on a single disc. This enables law firms, consultants and parties in the dispute to significantly reduce storage space and ultimately storage costs by imaging documents and decreasing the number of copies maintained. Team members in different geographical locations can cost effectively obtain an electronic image of all project documents without expensive copying costs.

- *Simultaneous Access to Documents.* Current technology enables multiple users to retrieve images and data into their workstations simultaneously. This eliminates the problem of tracking down documents that may be in the possession of another member of the legal or technical team. In addition, users in different offices can retrieve documents that are maintained in a central location through a network connection. They do not need to transport the documents from one site to another.

Document Database Considerations

In addition to the more obvious benefits of using computerized database support, there are some subtle, or even hidden, benefits. These additional benefits include:

- *Higher Productivity.* The time-consuming, expensive, and monotonous task of gathering relevant documents is drastically lessened. The document selection can be reduced to a few minutes of computer time spent searching while the sorting and printing of documents is left to the system to process.

- *Portability.* In the past, preparing for an out-of-town deposition, arbitration hearing, or trial meant shipping boxes filled with documents to a remote location. Lawyers would be accompanied by a full support staff to assist with document organization, identification and retrieval. Now, members of the legal team can travel with only a laptop computer and portable printer, accessing data by remote link.

- *Responsiveness.* Legal professionals must be prepared to use all available tools to access information as quickly as possible to counter arguments made by opposing counsel while in the courtroom, meeting room, or negotiating table. Using databases and images enables the legal team to pinpoint a document in seconds rather than spend minutes or hours manually searching for it.

Database Terminology

The following is a brief list of definitions of frequently used database terms:

- *Artificial Intelligence* refers to the capability of database search engines to locate relevant documents based on associated word groups or "concepts" rather than querying on keywords. For example, important documents related to delay may not contain the word "delay" but can be found using Artificial Intelligence searches for phrases or concepts associated with delays.
- *Coding* is the manual or electronic process of capturing key information from documents and transferring that information into database fields.
- *Database* is a collection of related information in an easily accessible electronic format.
- *Field* is a category of information (such as subject, author, recipient, date, OCR text, etc.) associated with each document record.
- *Hit* refers to a database record found as a result of a search query.
- *OCR (or Optical Character Recognition)* is an automated process in which scanned text images are converted into electronic text characters.
- *Record* is the coded field and OCR text contents captured for a single document.
- *Query* is a request to the database to search for records containing specified field criteria.

Document Database Considerations

DATABASE ASSUMPTIONS AND CONSIDERATIONS

Two important considerations in determining whether or not to automate a case include: (1) how much it will cost; and (2) how long it will take. The cost and time needed to develop a database, however, depend on various choices. For example, is it necessary to have all or only a portion of the document population contained in a database? Will the documents be imaged and those images linked to database fields? Are full text searches and retrieval of documents enabled by Optical Character Recognition (OCR) software required?

Automating the case will cost more money up-front than handling documents manually. The cost savings of a database over the life of the case, however, are significant because finding and retrieving data from an computerized database system takes much less time. Key documents also are frequently found using database searches that would most likely not be found using manual searches.

When planning a computerized database, it is important to identify the end results desired before determining the various tasks required to create the database. In order to determine the preliminary costs, answers to the following questions must be developed:

1) How many hard-copy pages of documents relevant to the dispute will be included in the database? What types of documents will be included in the database? For example, is it necessary to include hard copies of Primavera schedules and cost reports in the database? These documents, if determined to be useful in the database, will increase the overall database population estimate.
2) What percentage of the documents will be oversized architectural and/or engineering drawings? Is there a need to include oversized drawings in the database?
3) Will E-mail and other electronic documents also be produced electronically for inclusion in the database?
4) How many additional documents may be produced by opposing counsel during discovery? Have the parties agreed to produce documents for discovery in an electronic format?
5) Will additional documents be identified for scanning and inclusion in the database after further review of the issues and needs of the case?
6) How should the documents be organized and unitized prior to scanning? What image format and scanning resolution is required to optimize optical character recognition?
7) Should duplicate documents in the overall document population be culled, since the same documents are likely to be in multiple files or may also be produced by opposing counsel during document production? Some duplicate documents may contain marginalia which also may be highly relevant.

Document Database Considerations

8) Will all documents be imaged and electronically Bates numbered by a database vendor?

9) How will document exhibits be referenced during depositions? How will citations and exhibits be referenced in Briefs, Witness Statements and Expert Reports?

10) Should all scanned images be converted into electronic text using Optical Character Recognition (OCR) software?

11) Will the database include foreign language documents which need to be translated?

12) What database fields will be necessary? How will the documents be coded?

13) Which team personnel will have primary responsibility for checking the accuracy of the data fields in the database?

14) Which legal and technical personnel will need access to the database?

Document Database Considerations

RECOMMENDED PROCEDURES FOR PREPARING THE DATABASE

Database Development Steps

Computerized document databases are typically developed in several steps. Step 1 involves, where practical, organizing the documents by type and/or chronological order. Step 2 involves scanning hard-copy documents. Step 3 involves electronically branding each document image with a unique Bates number. Step 4 utilizes OCR software to convert the scanned document images into electronic text characters. Step 5 involves manually or auto-coding the documents and linking the coded data with the corresponding document images and OCR text. Step 6 involves loading the database into an in-house or on-line database repository. Step 7 involves database quality control and maintenance. Certain database vendors can perform some of the above steps simultaneously.

Step 1: Document Gathering and Organization

Depending on the size and type of document database, it is often useful to spend a limited amount of time organizing the collected documents prior to scanning and Bates numbering. Where practical, similar types of documents should be grouped together and placed in sequential or chronological order. Once the document images are scanned and Bates numbered the organization of the database documents cannot easily be changed. Therefore, a limited initial effort to organize the documents can often greatly facilitate the ease of working with database documents over the life of the project.

Step 2: Document Scanning

The technical and legal teams must determine whether all or only a portion of the project documents should be scanned. There are basically two choices available: (1) scan all documents; or (2) scan only relevant documents. It should be noted that when the decision is made to scan only a portion of the total population, attorney and/or expert review time needs to be incorporated into the budget figures to account for the process of deciding which documents are to be scanned. While the preliminary costs involved in choosing which documents will be scanned may appear high, the knowledge gained by the experts and/or attorneys reviewing the documents is maintained and used throughout the life of the case.

Choices Associated with Stage 1:

1.01 Scan all pages

or

1.02 Review documents and scan relevant pages

Database vendors typically recommend scanning documents with a resolution of at least 300 dpi using a black and white TIF file format to ensure high quality OCR. Half tone or gray scale settings should be turned off.

Step 3: Electronic Bates Numbering

For most cases, document images should be electronically stamped or "branded" with a unique sequential Bates number to facilitate document references. Often, this step can be performed simultaneously with scanning the document. It may also be necessary to reduce the document image to 98% its original size to allow space for the Bates number in the lower right corner of the image.

Step 4: Optical Character Recognition

Optical Character Recognition (OCR) is highly recommended for documents that contain a uniform typewritten format (e.g., correspondence, meeting minutes, etc.). OCR, however, will not accurately capture the content of handwritten notes or those documents that are not uniform (e.g., field notes, diaries, etc.). Based on Long International's experience in this area, approximately 75-80% of construction project documents contain the parameters required for OCR. Therefore, it is often more time-efficient to have a vendor OCR all documents rather than to sporadically choose a more limited subset of documents those that should be included in the OCR process.

Choices Associated with Stage 4:

2.01 OCR all pages

or

2.02 OCR a subset of pages

Step 5: Document Coding

Project documents should be coded to ensure the complete capture of factual information in preparation for the analysis of issues as well as for document production to opposing counsel. Objective coding of bibliographic, objective data is often best performed by an outside vendor. Certain vendors can perform this with an auto-code feature in their software. Such auto-coding has various degrees of accuracy and should be verified. However, this method will save time and costs. Subjective coding of legal and technical issues should be performed by client personnel, legal counsel and/or Long International personnel. This subjective coding may also be accomplished by performing searches and tagging documents into folders associated with issues. Keywords within documents can be found with searches and auto-coded into fields. Appendix A presents the potential code fields recommended to be included in a database. However, these fields may vary depending on the selected database vendor.

Choices Associated with Stage 5:

3.01 Objective coding of all documents

or

3.02 Objective coding of relevant documents

and

3.03 Subjective coding of relevant documents

or

3.04 Subjective coding of only key documents

or

3.05 Auto-coding of all documents

Step 6: Selecting a Database Repository

The use of an Internet-based database repository system provides the necessary tools to automate the search and retrieval of documents. With the support of an independent Application Service Provider (ASP), Long International can develop and implement a system that includes necessary design and security controls. Access is obtained via the Internet, allowing users in any location to access and search for documents. Changes made to any fields or information would be available instantly to all users. The ASP can provide the support necessary to train client personnel and legal counsel in the use of the software, creation and development of reports and segregation of records. The ASP also can provide telephone assistance 24 hours per day, 7 days per week.

Choices Associated with Stage 6:

4.01 Selection of the ASP vendor or an in-house database repository

Step 7: Quality Control

Depending on how the database was coded, certain quality control checks and corrections should be performed to enhance the accuracy of the database. This work should be performed by a dedicated team as well as by users as they see incorrect information while using the database.

Document Database Considerations

CONCLUSION

Document database services are volume driven. If there is either an increase or decrease in the volume of documentation, it will directly impact upon the database costs. Therefore, the volume of all documents anticipated to be received, including documents produced by opposing counsel during the discovery phase, should be considered in estimating the size and cost of the database. Once the database requirements have been established, vendors can be contacted to obtain pricing proposals.

Because arbitrations and litigations are typically on a fast track for hearings, all work associated with the case must be performed quickly and efficiently. Computerized document databases provide significant economic and strategic advantages. The legal and technical team can spend more time on analysis rather than searching for relevant documents. Most importantly, however, computerized document databases facilitate finding key documents which most likely would never have been located using manual searches.

APPENDIX A – POTENTIAL CODING FIELDS

Field No.	*Objective or Subj.*	*Field Name*	*Field Description*
1	O	BEGBATES	Beginning Bates number
2	O	ENDBATES	End Bates number
3	O	BEGATTCH	Beginning Bates number of attachment
4	O	ENDATTCH	End Bates number of attachment
5	O	TYPE	Document type
6	O	DOCNO	Document number (letter number, drawing number, etc.)
7	O	DATE	Document date
8	O	AUTH	Author and Author company
9	O	ADDR	Addressee and Addressee company
10	O	COPYEE	Names of individuals copied through cc or bcc
11	O	SUBJECT	Subject (re: line)
12	O	OCR	Text of document
13	S	ISSUE	Issue codes
14	S	KEYWORD	Keywords
15	S	PRIORITY	Priority (high, medium, low)
16	S	COMMENTS	Comments
17	S	PRIVILEGE	Privileged document (yes/no)
18	S	PRODUCE	Document produced for discovery (yes/no)

- (O) Objective bibliographic coding fields will be completed by a vendor.
- (S) Subjective coding fields will be completed by client, experts and/or attorneys.