Processing Basics

With the new revised Federal Rule for Civil Procedure already in place, it is a good idea to include some material here for new users to electronic discovery processing. Included is an overview on Electronic Discovery processing as well as a detailed list of file extensions.

Loading- This comprises of the loading of the data into a processing program. Loading assorted files is straightforward. However, loading emails can be tricky for first time users. If dealing with a common platform such as Exchange or Lotus Notes the emails are in what is known as a mail store. This is a storage device where email is placed. In email systems, each user has a private mailbox. When the user receives email, the mail system automatically puts it in the mailbox. The mail system allows you to scan mail that is in your mailbox, copy it to a file, delete it, print it, or forward it to another user. The mailbox format used by Microsoft Exchange® email systems is PST, while Lotus Notes® uses NSF files. Most programs will give the user the options if they want to include the attachments. If referring to emails only, they are known as the Parent and the attachments are known as the Children.

Metadata- Most programs during the loading phase will also have the function to extract Metadata. This is data about the data. It includes information about a document managed within an application or an environment. There are a few different types of metadata involved. It could be metadata based on a file system, document, email or customer added.

- File system- Data that is obtained or extracted about a file from the system storing the file.
- Document- Data stored in a document. Often the data is not immediately viewable in the software program where it was created. The best way to view this data is thru the "Properties" view.
- File System- Data that can be obtained or extracted about a file from the file system storing the file.
- Email- Data stored in the email about the email. This data is not even viewable in the application. The amount of email metadata depends on the email system.
- Customer Added- Data created by a user while reviewing the document. For example formulas created or annotations made.

Metadata should be extracted before any processing is done. It is not recommended to open up original native files. Why? If this is done then the metadata will be changed, usually the last modified date and possibly author. This could lead to spoliation sanctions. It is recommended at the very least that you make a copy of the source files.

Culling- This is the first phase of reducing the data set. This usually implies understanding what files can be processed and what files cannot. All system files usually are not processed, unless some of those files are dealing with an intellectual property case. Audio and video files are not processed either. These files are marked

as exceptions and included in a report. This is extremely important because from a legally defensible position every file must be accounted for.

Searching- Another way of reducing the data set is by searching. This is usually done via keywords. Searching can also be done by date range, or some level of file filtering. For those new to the world of searching via keywords, the most common is **Boolean**. An "and" operator between two words results in a search for documents containing both words. An "or" operator between two words creates a search for documents containing either of the words. A "not" operator between two words creates a search result for example containing the first word but excluding the second. It is imperative that those doing any type of searching know the basics. Otherwise the search hits may be invalid. Other search types can be proximity. For example "Lunch" within 5 words of "Box". Synonym searching is also another option. There are even more possible ways of search documents like phrase searching and sound a-alike searches.

The search techniques employed by the software is of the utmost importance. If the search engine(s) in the program are not robust it would be wise to use a search program to tag the hits such as DTSearch or comparable.

De-Duplication- This is the process of identifying (or in some cases removing duplicate documents.) There are different types of de-duplication as well. Here are the two most common.

- Project Wide- This is the culling of a document if multiple copies of that document reside within the same project. It ensures that only one of these documents is produced.
- Custodian- This culls the document if multiple copies reside in the same custodian's data set. This is the most popular way of doing de-duplication.

It is also imperative to understand the difference between identification versus elimination. Elimination usually does not leave any type of report. It is recommended to do identification so it gets logged. De-duplication is the most challenging part of the processes to understand. If done wrong by a user the whole project could be invalid.

Once the culling and filtering is done, the next step is the actual processing. There are a few options to think about.

- Paper- Take the culled down data set and print it to paper. It is the least likely to be used format, with the new rules going into place, but still an option. To some it is the easiest and most familiar to them. The negatives are costs for printing to paper and the amount of time it takes. Remember some excel and text files could be thousands of pages.
- Native- In the coming months there will be a rise in those wanting native productions. There are pluses to this including the cost will be significantly cheaper and the ability to view the document as the author

created it. A negative aspect is a high risk of spoliation. There are also issues of authenticity. There is also the inability to redact and assign bates numbers for native files.

- Image- This can be single and multi-page tiff, PDF. This is the most popular so far. The ability to extract the text and load it into a litigation database program, bates number and redact makes it enticing. The downfalls are it is expensive. Slowly the industry is moving away from the per page price and into the more reasonable per gig pricing.
- Combination- This is processing to tiff and producing in native certain file types. Some file types such as Excel are better suited for native production.

Processing to TIFF- In some cases processing to TIFF would be the next step. If doing a native production, the next step would be load file creation and some quality check before burning to a medium.

There may be some pitfalls on the front end when processing is started. I will go over a few of those.

- Password Protected Documents- If the documents are password protected then they cannot be converted. There are some options here. If the client supplied any passwords there are electronic discovery programs where you can enter those passwords. Where others do not allow for certain file type passwords. It could be a tedious process typing in a password as each document comes up. Another option is password cracking. Without the proper software it is recommended that the cracking be outsourced to professionals. The questions that need to be asked to the client are should the documents that were cracked be saved as a new file? If they are not then the metadata will be altered on the originals.
- Corrupted Files- Software knowledge is important here to try and fix those documents. The best option may be to mark those documents as exceptions and move on. Again, constant communication with the client is important to what they want done with these files.

The bigger a document means the longer it will take to process. Some large Excel files could take as much as half an hour to process. Understanding format options is also important. This can also significantly slow down the process. Have an understanding of the processing software you have. Know what the averages are per hour, its strengths and weaknesses. If it is a large data set this step could take the longest.

Quality Control- After the processing is finished the likely next step is some sort of quality check of the images. QC images is the process of reviewing tiff images for unsatisfactory output, deleting output on either a document or page level, re-printing tiff images to replace unsatisfactory output and creating custom message pages for documents which have no valid output.

Most off the shelf programs have a quality control module. Yet because of human error there is a chance a bad tiff will be missed.
The goals of quality control are to assure that everything is converted to Tiff correctly, checking for blank pages, making sure the text extracted worked, addressing any problem files and to make sure if a large file was encountered, that all pages were correctly converted to tiff. It allows the user to re-tiff documents that came back the first time as problems. Even reassigning a different document type to try and reprocess it. Also allows the user to do an automated cycle to check each page. However this is an imperfect science. It is foreseeable that the industry will be moving away from this type of manual quality check.

Production- After quality check the data is ready to be exported out of the program. It is important here to have an understanding of what a client wants.

- Single page or multi-page tiffs? PDF's?
- Are any file types being produced as Native?
- Do they want bates number burned into the image?
- Native files?
- Text Extract files?
- What type of load files?
- Custom metadata fields?
- Did you preload?

It is inherent to understand the different load files. Concordance use a DAT or TXT and an OPT file types. If the client works with IPRO as their image viewer then the file type could be a DAT and an LFP. Summation utilizes DII or SMI files. Ringtail uses MDB file types. It is always a great idea to preload the data into the program the client requests.

This article is nothing more than a bare bones look at Electronic Discovery processing. There are so many more things to know and understand that is not covered here. It is a start to understanding certain terms and what they mean. Best of luck!

Do not forget to do a virus scan before processing any data!