OCR vs. Coding – "The Myth and the Truth"

OCR – the great cure all,
It slices, it dices, and it makes great Julian fries,
It replaces coding - it replaces summaries,
It does everything you need,
And that's not all; you can go on vacation,
It's magic, right?

Well at least that's what some vendors and other purveyors of OCR software will tell you. Lets look at the realities of OCR and separate the myths from the truth.

For those of you not as familiar with OCR, it stands for Optical Character Recognition and represents the process where characters and numbers are converted from electronic image to electronic data (also referred to as optical scanning). An OCR software application reads the black and white pixels on an image and attempts to recognize the correct alpha character or numeric number, where they reside.

The result is commonly known as "dirty" OCR. It is called "dirty" as OCR will rarely be able to read even a pristine laser printed document with 100% accuracy. The worse the original - so follows the OCR. Guess how much accuracy your likely to get from a multiple generated copy? From the 1st to the second 2nd generation and on down, degradation enters the picture. Degradation is the process where not as many black pixels transfer, leaving letters that are not complete or where too many transfer, connecting letters in words. Remember, the OCR application uses the black pixels in its attempt to interpret the correct letter or number.

Fonts can create a problem as well. Look at the Old English style of lettering and you will quickly see why. Characters that touch, underlines and stray markings, especially those that cross text, also cause problems. Wherever OCR cannot distinguish the intended alpha or numeric character, it will often leave a slug or nothing at all. The characters or numbers that the OCR could not read are commonly known as slugs. Slugs are those non-alpha, non-numeric characters that usually show up as dashes, tildes, smiley faces and a variety of other similar items.

In addition, your original OCR is likely to have lots of spaces in between some words and yet other words in a sentence will all run together. This is due to the fact that OCR is neither perfect in its interpretation of alpha/numeric characters, nor whether a space exists between words. When OCR is attempting to interpret the words on an image that was created with full paragraph justification, the original document may have squeezed the words closer to together. This may result in the OCR application interpreting that no space exists between some words. This results in long strings of words running together.

But some will say - my application has a voting system, where it uses three different OCR packages to interpret the correct word or spacing intended. While this often results in better OCR interpretation, it still is not perfect and the same occurrences as noted above will result, just perhaps, less often.

Let's talk briefly about OCR cleanup. It is very tedious and expensive work and requires you look at the image or paper document in order to ensure you are making the proper correction. I once had an engineer doing OCR cleanup work while he was temporarily unemployed. After all, the case was a patent infringement and he would recognize the words and correct spelling better than the average person – so we thought. When we started to QC his work (yes, you have to do some sort of QC, even if only cursory), we noticed that entire words and sentences did not match. When we inquired of the individual, who was a PhD by the way, he simply stated he was writing the subject better than the original, so they could understand it better. OUTA HERE BUD!

Generally vendors will charge by the hour for cleanup work, as the realm of what they will have to do is so uncertain. They can't really tell if there will be a lot of errors or not. Often the subject will slow the correction process if many difficult terms exist, as typically seen in construction or scientific projects. If you need perfect OCR, you may be better off to find a keying vendor, who will double key, guaranteeing 99% or better accuracy.

That doesn't mean that we shouldn't OCR or use it as a research tool. In fact, it can be quite valuable in getting up and running quickly while you wait on a coded index to be completed. And you can't argue with the price, it's pretty cheap by comparison, generally $0.10 per page or less these days and I've seen it at $0.08 and $0.05 in limited cases. All OCR'd words are available and fully searchable, giving you full-text capabilities. OCR can often help when issues change in a case. I know that never happens in your firm, but I've been told it happens.

The main benefits to OCR are:

- Full text search capabilities
- Generally very inexpensive
- Get you up and running quickly
- Provides an inexpensive alternative to Keyword and Mention Names coding.

So are you ready to trust your database to just OCR?

If so, how, pray tell, will you do a chronology of the documents when you need it? When depositions approach and you want to try to pinpoint specific documents for a specific person or event, how will you do that with just OCR?

You cannot!

The point being made here - is that to use OCR as your only database tool, as suggested by some vendors or software promoters, is a Myth.

The Truth is, you really need to index/code the documents and capture at least some of the "Bibliographic" or objective information from each document into a database as well. (Note: Indexing and Coding are interchangeable words as used here).

Database Design and Considerations

Coding is the process of capturing information from each document into formatted fields in an organized and consistent manner. An organized and consistent manner is the challenge, but a must, if you expect to rely on the database, which acts as a fully searchable, computerized filing cabinet. Holding meetings with all necessary parties to discuss and decide the field names, formats, document types and specific capture instructions will help ensure consistency is maintained. When appropriate, make sample copies of specific or unique document types. If you do not maintain consistency, the database will not be as useful as it should be, or worse, not be trusted by the attorney(s).

NOW – let's remember that the document type field should be as broad as possible and is merely used to categorize documents. Don't get fancy here and try to use the doc type field for issues or topics, which are better stored in other fields. Build the database as a reliable work-tool for research. Don't build it or use it as an attempt to try your case with the database. Attorneys try cases using strategies based upon the documents found as they apply in theories of law or to the convincing of others. Databases store notes and information for fast retrieval and review.

Objective vs Subjective information. Objective capture requires no or little (if any) thought. Subjective capture requires thought, (reading, review and notations or summaries).

Objective or Bibliographic fields represent information that resides outside of the main body or text of the document. A person capturing information for these fields should not have to read or review the text or body of the document. These fields for example, would include:

- Beginning and ending document id number ("Bates" number)
- Date
- Author
- Recipient
- Copyee
- Title, subject or reline (if any)
- Characteristics (confidential, marginalia, etc) – Often optional
- Document type (doctype, as it is affectionately known, is somewhat of a subjective field, as it requires some thought process, though always included in objective-bibliographic coding)
- Beginning and ending attachment is also, often included. (Remember this field often depicts the physical binding elements, not whether it's the actual or intended attachment. (For example, a fax, letter and report, all stapled together). If your documents are being imaged, this will need to be captured at scanning).

A doctype list should be created that is specific to the collection. Keep the list short; remember general useful categories, not specific to each nuance. For example, if the financials are not really part of the case, why not place all tax, invoice, purchase order, P&L, balance sheet, 10K, etc, type documents into one doctype FINANCIAL. If they are important, decide which ones need to be specific. For example, if the case involves tax filings, then maybe split those out.

Remember this helpful hint. If you are also capturing titles, there is a "sort of" secondary type available through the title. For example, you search for the doctype FINANCIAL and the title, where the title includes 10K. You get a list of hits you can now review and add notes or complete issues or topics. You can see that the single doctype FINANCIAL is more than sufficient here when used in conjunction with other fields. The less a coder has to remember or interpret, the better. Otherwise, imagine 20 coders, all guessing, at the category or checking off the first one that appears correct, when one further down the list of 63 was more appropriate. You get the picture.

A document type list should be short 12-20 doctypes should do. Don't try to issue or topic code with this field, you may regret it. You can use the same idea on correspondence, as apposed to fax, letter, memo, etc, or many other fields. It all depends on your case needs.

DO NOT include punctuation in the title field or any field, except for example, as a separator between multiple names. No one ever searches for punctuation. Remember you are building a database for research, not looks. And never, ever use a double quote (") sign anywhere. It is generally used as a separator in providing comma-delimited ASCII loads. If you need to see the punctuation, then look at the image.

Write all rules for each field into a Coding Manual or set of Coding Instructions for approval by the litigation team members and review the manual with the coding team before you start coding documents. There are many items, to numerous to cover in detail here, which affect the consistency and format of your data.

For example, specifying last name first. Do you use a comma or a dash before the first name? Are you going to include spaces, initials, or Sr. or Jr. titles? If so, specify how they should appear and show examples. If you are going to include the associated organization when listed, how do want to see that? If your using JFS, be very careful how you specify the capture of data for names.

Another example worthy of note is date. Does it represent the actual origination date and if not, do you need a date-estimated flag? When there is clearly no origination date, do you use the earliest or latest date? There are many decisions to make, none more particularly correct over another, however a standard must be decided and maintained to ensure consistency. Like dust is an enemy to computers, inconsistent data is the enemy to databases. It will quickly become your enemy and nightmare if not adhered too. It is imperative that you include at least a few bibliographic fields to be able to sort hits in chronological order or other specifics as may be necessary in research during discovery and after.

Additional fields not typically associated with subjective coding would include Mentioned Name and Keywords. These require a simple cursory view of the text, but not an in-depth reading or understanding of the document. Remember, vendors usually price these services by the page and just one or the other is two times or greater than the price of OCR.

Subjective coding generally includes fields, which require the reading and comprehension of the subject of a document in order to correctly capture, topics, issues or to write summaries. These are expensive captures when one considers the time element to perform this task as opposed to bibliographic coding. If you are using a vendor, have them capture the bibliographic information and do the other coding at the firm with persons close to the matter.

Consistently captured bibliographic data will provide you an opportunity to successfully perform relational and/or proximity searches. Relational searches are used for example, to find a letter, by an author - to a recipient - between certain dates – regarding such and such. A proximity search is where you look for one word adjacent to another or within so many words of another word).

So where do we stand with the Myth and Truth?

As you can see, OCR should not be used alone without a coded database. A database, which consists of standardized coded data, can go a long way toward sifting and sorting information to find those golden nuggets or that smoking gun. For collections where OCR makes sense, the combination of OCR and coded data can be a very useful and powerful tool, allowing you to make more efficient use of your time.

Some brief comments on Cost Control Benefits of Automated Litigation Support.

Clients are demanding more accountability and efficiency from their legal counsel regarding the cost of litigation. The use of imaging, OCR and database building can play a major role in reducing cost during discovery and again at trial. Data must be sifted through and sorted into manageable blocks of information that can be quickly called upon as needed. The vast majority of data compiled will not be important to the pursuit of the case, but all must be evaluated at least once, and set aside in an orderly manner.

The use of imaging, OCR and coding, converts unwieldy paper files into manageable information. No more lost time digging through, sorting, copying and stacking and re-copying and re-stacking or re-filing documents. No more lost time looking for lost documents.

Mantlemen's Imaging Guide, 1st Edition, 1995 states that on average:

- Companies make 19 copies of each document
- Workgroups lose 15 % of all documents they handle
- 7.5% of all documents get completely lost.

The cost associated with filing, retrieving, duplicating, distributing, refilling and storing document are significantly reduced when using an automated litigation support system. Outsourcing the conversion of the files to a vendor can significantly reduce the cost of imaging and database building.

The following tasks are most economical when priced by the page; imaging, OCR, Mentioned Names and Keywords. Coding should be priced by the document as opposed to by the page, as only one set of bibliographic fields will be captured per document, regardless of the number of pages in the document. Coding bibliographic fields generally involves looking at only the first and last page of a document.

Review of the benefits gained:

- Portability
- Cost savings
- Increased productivity
- Accuracy and consistency ensured
- Avoid lost, misfiled, or out of file documents
- Reduced cost in imaging and database building

Of course, to obtain these benefits, one must use an experienced vendor. Be sure to interview at least two vendors and get references. Make sure they have the necessary experience to handle the complexity of what seems so simple. After all, anyone can create an image. Its what you do with it that counts. Its not like making copy sets.

Clients know that legal assistants and associates do not learn anything from a document by capturing its bibliographic data. Spending hours on such repetitive tasks can raise eyebrows at billing time. Once the data is captured via a more efficient, unit priced method, you can perform research and review of the document categories for addition of topics, issues, notes, comments and summaries, generally billable work. After all, do you really want to trust your issue coding or summaries to 20 coders, all who have no particular interest in the case or only to get potentially 20 different answers.

It is generally to everyone's advantage to consider the most effective way to manage the paper, microfilm, fiche, and emails in a lawsuit. While imaging and database building generally occur at the front of a case, the cost savings are significant when amortized over the life of the case. Amortization of the cost is a significant consideration in getting started early when paper hits the door.

Open Door Solutions, LLP is a Dallas-based company providing litigation support services and document management solutions to law firms and corporations nationwide. Author Bob Sweat received his education in Business Administration and Economics at the University of Wisconsin and advanced work at Purdue University. He holds a Paralegal Certificate in Civil Litigation with Computer Emphasis from The Center for Legal Technology, Milwaukee, WI, and has years of experience working with local and national vendors on large, complex litigations. Bob is currently a partner at Open Door Solutions.