



Metadata V. Coding in electronic discovery

There is a common misconception in some parts of this industry that an electronic discovery project does not need coding because it has sufficient metadata. What we need to remember is that metadata is just that the data of the document, or the properties, sometimes referred to as the profile of a document. Metadata does not reflect any information within the document and does not pertain to what is entailed in the actual document. Metadata could also be misleading as opposed to the content. For example, a document date in the metadata can be different from the content. Document metadata could have a creation date of 12/01/2005, where the actual date in the content could be 12/05/2005. Which one takes precedent? Which date is more accurate in helping a law firm run accurate searches? What if you need actual bibliographic data such as document type? This will not be in the metadata. What if someone edits a document, the metadata in the author field will not reflect the original author?

In the past, an auto-coding program had its share of limitations since it was working off of scanned documents. Auto-coding could only be as accurate as the OCR. However, in electronic documents the accuracy is close to 100% because there are no formatting or OCR issues to deal with. Because of this, an auto-coding program is more ideal to do bibliographic coding. This document will look at electronic discovery and auto-coding. One of the best programs on the market that customers can purchase is ALCoder. This program goes into native files, emails and attachments and extracts the objective information from the document, not the metadata.

History of ALCoder

From Rosen Technology Resources

Keith Rowand, the software architect behind ALCoder, began selling ALCoder in 2001. After several years of attempting to develop, market and sell the software, he partnered with Rosen Technology Resources. RTR is an award-winning litigation support consulting and training company. RTR has exclusive rights to market and sell the software and has taken it to the next level, developing

partnerships and integrations with software like Concordance, Summation, ICE, CaseMap, and iCONNECT. More software integrations are nearing completion in the coming months.

ALCoder auto-coding software has the ability to extract the following 14 fields of information from within a document: Date, Estimated Date, Document Type, Title, Title String, Author, Recipient, Copyee, Names Mentioned, Organizations Mentioned, Dates Mentioned, Numbers Mentioned, Keywords Mentioned, and a Document Score. The document score is based on several criteria including how many fields of information could actually be extracted from a document. For example, if a document is a handwritten note, it will likely get a score of an "F", whereas a memo would likely get a score of an "A." In addition to ALCoder auto-coding paper documents, it is also used for auto-coding ESI. In fact, ALCoder does a tremendous job of coding ESI because there are no formatting, white space or OCR issues to content with, as there are with paper documents. In truth, ALCoder should be used for all ESI because the metadata that is extracted during the processing of ESI is often not accurate or complete. The metadata can often reflect the owner of the software as an "author" instead of the actual person who created the document. The metadata "date" extracted may reflect when the data was moved from one server to another, rather than the actual date created. There are many factors that prove that it is not practical to rely on metadata as accurate bibliographic information.

For each "batch" of data that is processed through ALCoder, it can actually get smarter. ALCoder allows for editing of authority lists (the "AL" in ALCoder), and as the lists get edited they are used for subsequent batches. The authority list extracts People and Organizations. A user can import a list of names and organizations, let ALCoder extract the information and do its own edit, or manually edit. Keywords can also be imported into ALCoder and extracted to a keywords field. In the Workstation version or higher, ALCoder can also identify near documents based on a threshold that the user sets. ALFind is also an add-on to ALCoder that can be run in LAW or Concordance and populates a Similar Documents field with hyperlinks to all the versions of a document.

There are three versions of the software, and the high-end version is hardware dependent, so it can process as fast as the machines can process. On average, it can auto-code approximately 4-5,000 documents per hour, based on a five-page document.

ALCoder is also seamlessly integrated with Concordance, Summation, LAW and CaseMap for one-click coding. The software is purchased by law firms, government agencies and service bureaus, and the only auto-coding software that does not charge "click fees."

ConceptuAL was created in 2006 to supplement the limitations of database searching and allow for concept searching within Concordance and Summation. Simply put, concept searching provides more focused results when searching for relevant data, without all the "noise". When searching for a phrase in conceptuAL, conceptuAL is not just looking for the words or phrases, conceptuAL is identifying related concepts and phrases to expand the search to include information that relates to what you want without knowing all the ways to say the same thing. It also provides the ability to refine a search based upon those concepts.

For example, a search for the word “trunk” may return results for: 1) tree trunk, 2) car trunk, 3) elephant trunk, 4) shipping trunk. You could use conceptuAL to refine your search based on the definition which would significantly lower the number of documents to review. And to expand, you could do a search for the words Financial Agreement and while those exact words may not be in your database, the results returned would be similar based upon the concepts.

Our toolkit of ALCoder, ALFind and conceptuAL allows the user to Code, Dupe and Search without ever having to send the documents out.”

Examples of Metadata v. Coding

A few Microsoft Word, Excel and PDF documents from a 2006 data test set were run thru a few electronic discovery programs. The documents used were from an actual case and picked randomly. Filenames and authors have been renamed in these examples to ensure integrity of the original files. Text was extracted along with a metadata load file. A text file control list was created. ALCoder needs some sort of load file in order to bring in OCR and/or full text. Here are some of the findings:

#1

The first document tested was a Microsoft Word document. The filename was generic and called Jan2002. The contents of the document were in regards to a National Sales Meeting. The Author was extracted and for this exercise and to keep the identity of the author confidential, I will rename the author to Pronk01. The metadata listed as the Date Created as 1/10/2002 and Date Modified as 1/11/2002. The EDD program extracted the title as 30-Nov-01.

Now let us look at the coding that was processed by ALCoder from the text extracted. The document date listed was 1/11/2002 which is what is on the first line of the document. The document title was National Sales Meeting. ALCoder is able to even give a summary of the document. The document type is listed as an Article. ALCoder is able to give additional fields such as Mentions and Organizations. This helps with doing searches for other people and organizations that may be deemed worthy of responsive or privileged searches. Since there was no any mention of an Author in the contents of the document, ALCoder was unable to find an author.

Filename	Document Title
Jan2002 (EDD Program)	National Sales Meeting (ALCoder)
Title: 30-Nov-01	

Author: Pronk01 (EDD Program)	NA (ALCoder)
Date Created: 1/10/2002 (EDD Program)	Document Date 1/11/2002 (ALCoder)
Date Modified 1/11/2002	NA (ALCoder)
Subject: NA (EDD Program) NA (EDD Program)	Summary: Detailed overview of Document (ALCoder) Additional Fields of individual names, organizations and dates. (ALCoder)
MD5 Hash Value (EDD Program)	NA (ALCoder)
Source Application: Microsoft Word (EDD Program)	NA (ALCoder)

Analysis of document #1

The EDD program was unable to extract all metadata for this document. Last saved by, Company and date last printed was unable to be extracted. The question becomes whether a law firm that maintains this data in Concordance or Summation has enough information about the document. The title from the metadata says 30-Nov-01 and the filename is Jan2002, which means it is possible someone may have renamed this document. What is more important - the date the document was created or the document date in the contents? Would you rather have the filename or the actual document title? In terms of electronic discovery, does it make sense to only have Metadata? The coding from this automated program has given us the actual date on the document, the actual document title, a summary of the document, and additional mentions of individuals, organizations and dates. What this first example shows is the need to not only have metadata extracted from a document but some sort of coding done to find additional information about the actual contents of said document.

#2

The next example is of a document that contains a travel itinerary for an individual. It is also a Microsoft Word document. The metadata that was extracted was as follows: the filename is itinerary.doc, the title is Travel details for, the author is Smith, last saved by is Kwan, date created is 1/15/2002, date last modified is 1/18/2002, date last printed is 1/18/2002. Again, the question that needs to be asked is whether there is enough information from the metadata to do proper searches and determine if this document is responsive, non-responsive or privileged? And is the metadata accurate? A different EDD processing program was used to extract the metadata from this example. For some reason the last saved by information was not included. Let us now take a look at what coding was processed.

The Document Title is Travel Details for: Michael Ward, the document date is 1/21/2002. The coding also includes additional names, organizations and phone numbers. Once again, the question becomes which date is more important? Title? Having metadata and coding for electronic discovery seems more advisable.

Email

This is from a random e-mail from a PST. Let us look at the metadata extracted.

Author: Cynthia Price

Recipient: Stuart Redding

Subject: Re: Sample

Sent: 6/6/2006 10:01 EST

Date Received: 6/6/2006

Filename for attachment: Presentation_June_2006

The metadata also included the location of the PST, the date created and last modified. In this example, the coding produced a summary of the e-mail as well as the company name and contact information of the author. Overall, the metadata seemingly is more defined for e-mails than loose files. What these random tests have done is show that not all metadata is extracted from an EDD program. Additional tests were run with similar results.

Why should a service provider have an auto-coding program in-house?

With technology ever-evolving and education being so important, it becomes all about options. Do you really want to continue to outsource every single coding project? Even the small ones? What if you have a client that wants you to also give them a coding load file of electronic discovery files after it has been processed? Then the client would have to wait until you send it or FTP the files to your coding vendor. Then after it comes back it may still need some sort of quality control to make sure it looks correct. There may be some service providers that still do coding in-house. If that is the case, it would be great to have an auto-coding program handle the electronic discovery projects, at the very least. Does it really make sense to send electronic discovery files to a coding vendor when there is software available to handle it, along with staff that is educated and understands coding?

One of the big weaknesses of auto-coding software has been the issue of its accuracy. Since the majority of these programs rely on OCR, it has come down to more of a hybrid coding solution - auto-coding with quality control to manually code those documents that the program could not read because of bad OCR. With electronic documents, this is not the case. After the documents have had text extracted, they can then be loaded into ALCoder and processed to produce the specific fields necessary. The accuracy is close to 100% because they are electronic documents.

There are numerous schools of thought in dealing with electronic discovery for service providers. Some small shops have a few licenses of the less expensive electronic discovery processing programs and assume they can now provide electronic discovery. However, these providers lack proper training and understanding of EDD processing. Then there are a heap of tier 2 providers (those that provide electronic discovery, scanning and copying) that believe they only need an electronic discovery processing program in-house to compete. On a whole, these providers have a better understanding of electronic discovery, however they are limiting the options to their clients because everything must be run through the electronic discovery processing application. If a client wants electronic discovery documents coded, it often has to be sent out.

If a client wants one of these tier 2 providers to take a 2 gigabyte PST file and run a few searches and then export out the results into a new .PST file, how will these providers proceed? The majority of the processing programs cannot create a PST file. It truly comes down to one word - options. For a service provider to compete and be well-respected, it needs to provide electronic discovery options. This is where having an electronic discovery toolkit comes into play. Having a processing program is not enough; the provider also needs to have additional software. Is having software that covers the majority of the electronic discovery lifecycle enough? No, that is half the battle. The other half is education and having people on staff that understand electronic discovery and how these programs work. What is the best way to get the client what they want and to do it the right way each and every time?

Why should a law firm have an auto-coding program in-house?

For those law firms that send out coding for electronic discovery cases, it is not practical. It is far too expensive and time-intensive to do that for all projects. The ideal situation would be to have an auto-coding program in-house to process the electronic documents. Delegating a senior analyst on staff that

understands electronic discovery and can process documents with a program such as ALCoder would save the firm time and money.

Additional program from Rosen Technology Resources

ALFind

ALFind 3.0 is a utility that identifies near-duplicate documents in existing databases such as LexisNexis Concordance and LAW, and CTSummation.

ALFind identifies near-duplicate documents by comparing e-mails, native files, and/or OCR using a similarity threshold established by the user. ALFind can be used as a stand-alone application or can be run in existing databases.

"Near-duplicate documents often represent a large percentage of electronic document collections. Using ALFind, the case team can eliminate duplicate documents or review similar documents that have slight variations", says Lisa Rosen, President of Rosen Technology Resources.

Two versions of ALFind are available, both of which are available for a one-time cost, with no per-page, per-document or per-gigabyte fee.

For more information or to request a demo, please contact sales@rosentech.net or 312.251.4440.

Rosen Technology Resources also offers evaluation copies of ALCoder. To find out more about ALCoder and their toolkit please visit their website. <http://www.rosentech.net/>

For the last two years, Rosen Technology Resources has won the 2005 and 2006 Law Technology News Litigation Support Service/Consultant Award.

Randall Consulting highly recommends having an auto-coding program in an electronic discovery toolkit. ALCoder is the best program this author has seen for automated coding. This goes with having educated people on staff that fully understand electronic discovery and the difference between metadata and coding.

This paper was written by John Randall who has seven years of experience with litigation support and electronic discovery. This is the third of four free white papers in 2007 on programs that cover the litigation support life cycle. No compensation was accepted in writing this neutral article.

ALCoder will be talked about in more detail in regards to scanned documents and OCR in the electronic discovery toolkit book coming soon!!!!

Electronic Discovery processing and overview Boot Camp Classes coming soon!!!

The next Randall Report on analysis of electronic discovery processing programs will be released in the fall of 2008.

Any questions on this article please e-mail John Randall at jrandall@randallconsulting.net