CDS COMPLETE DISCOVERY SOURCE

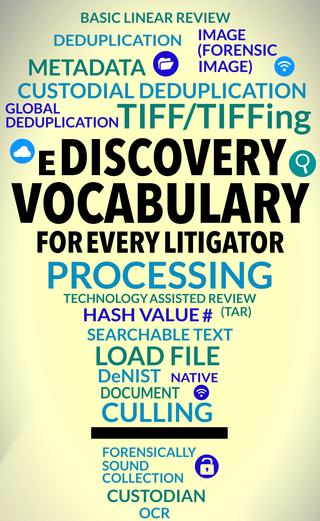
eDISCOVERY VOCABULARY FOR EVERY LITIGATOR

We have all heard that attorneys speak legalese, and we know of the "IT speak" that flies over the heads of IT industry outsiders. How can we manage the additional eDiscovery industry terminology that is a necessary part of the conversations between IT and legal that surround litigation or a government investigation?

Here's an eDiscovery Vocabulary List. Keep in mind that these definitions are intentionally broad, intended to give a basic understanding that may just barely scratch the surface for some of these terms.

CULLING:

A broad term that is simply the act of removing documents from a collection in an attempt to reduce the size of the collection. Some standard ways to cull are DeNIST, deduplication, applying date ranges, running search terms, and some forms of analytics.



DENIST:
The National Institute of Standards and Technology ("NIST") has a running list of non-user generated document signatures that it has established as having little or no value for litigation purposes – in other words, it is an industry accepted list of "junk" files (mostly program and system files that do not contain user-generated data). When you "DeNIST" a collection of ESI you are simply removing these industry accepted junk files from the collection.

DEDUPLICATION:
Deduplication: the process of removing duplicate files from a collection of ESI based on their hash values.

GLOBAL VS. CUSTODIAL DEDUPLICATION:
Two different ways of running deduplication. Custodial Deduplication removes all duplicate files within a single custodian's collection. Global Deduplication removes all duplicates across all custodians in a matter.

GLOBAL DEDUPLICATION:
If the collection is globally deduped, Custodian A will end up with one copy of the email, but both versions of the email will be removed from Custodian B's collection because Custodian A was considered a higher priority custodian and only one version of a document can exist when globally deduping.

CUSTODIAL DEDUPLICATION:
If we dedupe this collection by custodian, Custodian A will end up with one copy of the email and Custodian B will also end up with one copy of the email.

DATA TYPES:

DOCUMENT:
A specific file, i.e. email or Word document. Sometimes a specific file ("Parent") can also have files within it ("Children") and that entire group of documents is called a "Family." Ex. an email with two attachments, the email is the Parent, the two attachments are Children, and all together are considered a Family.

NATIVE:
A document that is the format that it would naturally appear on your computer, such as a Microsoft Word or Excel document. This is opposed to an image format.

METADATA:
Self-created data within a file. It can be created to record various elements of the file, such as the name of the document or when it was created. In an email you would find such metadata as the time the email was sent, who sent it, who received it, and so on. Different files store different types of metadata.

SEARCHABLE TEXT:
The body of a document is made searchable when the text of the document is indexed. Just like an index in a book records pages on which various words appear, this index records where various words are located within a collection of documents.

eDISCOVERY TECH TERMS:

IMAGE (FORENSIC IMAGE):
A bit-by-bit copy of a computer's hard drive, which essentially equates to a full and exact copy of the entire computer. Once an image is taken, you can open it up, or "mount" it, to look at the image exactly the same way you would have looked at it on the original computer the moment the image was collected. This is the most inclusive, broad, and complete form of collections.

OCR:
Optical Character Recognition ("OCR") is a way to get searchable text from a document that does not have any. Essentially, this technology converts a picture or image of text to usable text by scanning over the image to identify a character and recording it as text.

TIFF/TIFFING:
TIFF/ TIFFing: TIFF (Tagged Image File Format) is simply an image format, like JPEG. What is generally referred to as "TIFFing" is the act of converting, or printing, a native file to this image format, much like when you take a Word document and "Print to PDF." Documents are TIFFed for production for much the same reason as you would print a document to PDF before sending it to someone, which is that it memorializes the document.

LOAD FILE:
Similar to an Excel spreadsheet, a database is neatly organized into rows (referred to as "documents"), and columns (referred to as "fields").

HASH VALUE:
A value that is automatically assigned to data based on an algorithm that assesses many aspects of the data. This algorithm is accepted by the industry to be so thorough that if two pieces of data have the same hash value, then they are considered identical.

PROCESSES:

BASIC LINEAR REVIEW:
Reviewing documents one after another as they naturally appear in a collection.

CUSTODIAN:
The person or entity that is responsible for, has administrative control over, or has access to electronically stored information ("ESI") – basically it is who "owns" a document.

TECHNOLOGY ASSISTED REVIEW:
The broad concept of using technology to organize or expedite review. The term is most frequently associated with Predictive Coding, a specific type of analytics that will be discussed in a future article.

FORENSICALLY SOUND COLLECTION:
In most cases, a full forensic image is not necessary and a more "targeted" collection method is sufficient. If that is the case, any number of other collections methods may be used as long as they are "forensically sound." This term means that the collection happens in a manner that ensures the collected documents, including their metadata, are not altered in any way and the resulting collected documents are identical to the documents as they originally existed.

PROCESSING:
The process by which metadata and searchable text are extracted out of a Native file and put into a usable (i.e. searchable) format. This metadata and text is also what is analyzed when analytics are run on documents, which will be covered in a future article. Deduplication and DeNISTing of the collection often also occur during processing.