Published on *The National Law Review* https://natlawreview.com

What is Document Processing? And What Factors Should You Consider When Processing Documents?

Article By:

John C. Molluzzo, Jr.

While document collection, analytics, and review are crucial—and much discussed—phases of e-Discovery, document processing is similarly important and can have lasting consequences for any review of electronic data. Document processing consists of taking the raw data collected from its source and transforming that into a readable format that can be used and manipulated by attorneys.

Document processing includes, among other things:

1. Extraction: Extraction involves unpacking documents such that native files are separated into separate documents. For example, a single email with an attachment will become two records, a parent email and a child attachment. Similarly, as discussed below in further detail, files embedded in other files (for example, a spreadsheet in a PowerPoint) may also be extracted and create a parent/child relationship.

2. Metadata Processing: During document processing, metadata will be extracted, normalized, and preserved. Typically, metadata will then be placed into fields, which will be viewable in the review platform used to review the documents. At this stage, documents are also typically assigned a unique identifier (often called a Document ID or Control Number), which allows each document to be identified and tracked.

3. OCR (Optical Character Recognition): Documents will typically be subjected to an OCR process, which involves taking images of non-searchable documents and identifying text in them, so that they are searchable. Not all files will be searchable, and thus, if your document set includes unique file types, be sure to consult with your e-Discovery team to identify whether such documents can be OCRed.

Additionally, other factors that one should consider when processing documents include, but are not limited to:

1. Choosing a Processing Time Zone: When processing one must choose a time zone to be applied across the documents. While this may seem like a choice of little import at the beginning of a case, it will govern how documents appear when they are reviewed and

produced. What time zone to choose will be governed by factors unique to each situation—for example, the location of the documents, where a client is located, or perhaps the Court in which a case will be litigated.

2. Whether to Extract Embedded Objects: Embedded objects are files contained within other files. These often include PowerPoint files that contain embedded Excel spreadsheets. If these files are extracted, they will be separated into separate files for purposes of review. For example, if extracted, a PowerPoint file with 3 embedded spreadsheets will become 4 documents after it is processed—the original PowerPoint and three separate Excel files. This may be useful, for example, in a case where the underlying spreadsheets relate to the key issues in the case and contain important information. Extracting these spreadsheets as separate Excel files allows the reviewer to examine them more closely. Of course, extracting embedded objects like these will increase the total number of documents for review (sometimes significantly), and thus can significantly increase the pool of documents for review. In contrast, if these embedded files are irrelevant to the litigation (as is oftentimes the case), there may be little reason to extract them, as the spreadsheets remain viewable as charts within the PowerPoint itself. Each case has unique needs, and one should discuss this issue with the document processor to ensure that documents are processed in accordance with the needs of a particular situation.

3. Whether to Extract Embedded Images: Certain embedded images can be ubiquitous—for example, logos that are found in email signatures. Typically, these file types are not extracted during processing, unless a specific case need requires doing so.

4. Deduplication: Counsel should strongly consider whether to use deduplication, which typically uses unique hash files (often called the MD5 hash value) assigned to files to remove [de]duplicate documents during processing. Deduplication often increases efficiency by reducing the number of documents that must be searched and reviewed.[1]

In addition to the factors above, when document processing is underway, it is usually helpful to provide the e-Discovery vendor or document processor with other specifications that will be applied to the data source—for example, applicable date ranges and other limitations, such as the names of custodians to be searched. Finally, once documents have been processed, one may apply search terms to further narrow the data before it is reviewed.

Rules vary from jurisdiction to jurisdiction (and sometimes from judge to judge) regarding how much information parties are required to share with one another regarding e-Discovery, including processing specifications. Some require more extensive disclosures, whereas others require none. Nonetheless, regardless of what is required, it is often useful for parties to meet and confer at an early stage of litigation to exchange information with regard to e-Discovery, including how documents will be processed. In doing so, attorneys can discuss any issues at the outset, with an eye toward resolving any conflicts before the parties have spent significant time and cost on the discovery process.

[1] Another mechanism sometimes used to reduce the number of documents in need of review and to maximize efficiency is threading. To learn more about threading see this post by Litigation Shareholder Katy Cole.

National Law Review, Volume XIII, Number 17

Source URL:https://natlawreview.com/article/what-document-processing-and-what-factors-should-you-consider-when-processing