Terminology:

Electronic discovery (e-discovery) refers to discovery in which deals with information in electronic format, or electronically stored information (ESI). Electronically stored information is different from paper information because its form is intangible. Examples of the types of data included in e-discovery are e-mail, word documents, PDF documents, spreadsheets and any other electronically stored information which could be relevant evidence in a law suit. In addition, electronic information is usually accompanied by metadata, which is not present in paper documents.

Metadata is data about the electronically formatted file, or data about data. Examples include creation/modification date, creator/modifier, text, and the application that created the file. The preservation of metadata from electronic documents creates special challenges to prevent spoliation. Metadata is sometimes relevant and can play an important role as evidence in litigation.

Native format is the original file format. The information is viewed in the original application that created the file. Native format is editable and subject to spoliation.

Render or Image format is essentially a picture of each page of the original file stored in tiff format. The document is viewable with an image viewer. Metadata from the native format is extracted and stored in an external storage point. Corresponding text from is stored in an associated text file. The data in the text file is created either from metadata extraction or OCR processing from the image. Image format is the preferred format for Summation.

Petrification involves the conversion of native files into an image format that does not require use of the native applications. This is useful in the redaction of privileged or sensitive information, since redaction tools for images are traditionally more mature, and easier to apply on uniform image types.

Traditionally, electronic discovery vendors such as DTI and ELS have been contracted to convert native files and paper documents into TIFF images (for example 10 images for a 10 page Microsoft Word document) with a load file (DII) for use in image-based discovery review database applications such as CT Summation. Increasingly, database review applications have embedded native file viewers with tiff capabilities. With both native and image file capabilities, it could either increase or decrease the total necessary storage, since there may be multiple formats and files associated with each individual native file.

Data Streams are data in native format that need to be processed in the Preparation stage. The streams can be loose files or archived files (e.g., zip and pst). An assessment of the data size and time frame to be completed are required to establish the parameters to determine internal or external processing.

Target format is the type of output to create for the review process. Either native or image format is acceptable for Summation.

Native format are files produced in the format they were created and maintained. In a native production, MS Word documents are produced as .doc files, MS Excel files are produced as .xls files, and Adobe files are produced as .pdf files, etc. Native format is often recommended for files that were not created for printing such as spreadsheets and small databases. For some file types the native format may be the only way to adequately produce the documents.

For instance, Microsoft Excel spreadsheets do not lend themselves to being converted to image because the worksheets often do not conform to a standard 8 ½ by 11 inch page. Even if the number of rows and columns do conform to a standard size of paper, there are often formulae and other information that is essential to the matter at hand that requires that the files be produced in native format. Small databases are another good example of native data that may best be produced in native format.

E-mail is not typically produced in native format. E-mail is typically stored and maintained in an e-mail system that is like a database (MS Outlook, Lotus Notes, Groupwise, etc). E-mails may be exported to native format from the e-mail system in various formats including .psts for Outlook and .nsfs for Lotus Notes. These files are typically converted to individual files during processing for the document review. There are numerous e-mail systems and users may have various methods for saving or archiving e-mail. In these instances, some e-mails may be produced in native format. One example of a native e-mail production may be where a user saved individual e-mails outside of the e-mail system as Outlook .msg files. . These individual messages could be reviewed and produced in native format.

(http://edrm.net/resources/guidelines/edrm-framework-guides/production-guid#4-1-1-native-file-formats)

CT Summation allows for review and production of native file formats. In Summation 2.7 a new eDoc viewer was introduced that allows viewing of file formats without using the native applications. The new eDoc viewer allows the document reviewer to view the document without the potentially altering the document. In addition, the system now allows the ability to compare differences between near duplicate documents using a new sub system called "Win Merge".

Image format are files produced into images files, such as tiff or jpeg. Rendering an image is the process of converting a data stream or scanning paper into a non-editable digital file, such as

a PDF or TIFF. During this process a "picture" is taken of the file as it exists or would exist in paper format. Based on the print settings in the document, the printer or the computer, data can be altered or missing from the image. Expertise in the field of e-discovery and image rendering tools are necessary to minimize this risk. The table below lists some common issues when rendering data streams to image format or paper format.

File Type	Potential Risk
MS Word	Auto-dates may display the date the files were converted to image; comments may or may not be displayed; track changes may or may not be displayed; links may not be apparent
MS Excel	Hidden cells, rows and columns will not be displayed; comments may or may not be displayed; formulas will not be displayed; links may not be apparent
MS PowerPoint	Speaker notes do not print by default; animations will not display properly; links may not be apparent
Various	Embedded images in original not translated into image version and may not be visible

Search criteria are parameters that are used to review the data stream and select in bulk files to prepare for review. The parameters are agreed upon statements to evaluate the data for high level relevance to the case. These may include creation/modification time frame, creator/modifier, file type, keywords in the text and other metadata filter criterion. Search criteria are vital in the reduction of data stream size and processing time during the preparation stage.

Hashing consists of running a file or other piece of data through an algorithm, creating a unique value for every possible piece of data. This unique value is known as a hash value. Running a file through such an algorithm ensures that only exact duplicates will be identified.