fieldfisher

eDisclosure Glossary

The world of eDisclosure can be technical and complex. Our Dispute Resolution Team has created a high-level glossary to help parties engaged in eDisclosure understand some of the basic terminology used.

Algorithm (for Technology Assisted Review)	A detailed formula or set of steps used to solve a particular problem, i.e. sorting documents as relevant / not relevant. A "seed set" is used as a representative cross section of a data-set and documents are trained by an expert reviewer as either relevant or not relevant. An algorithm is generated over the seed set and this is used to predict the relevance status of future unreviewed documents.
Basic Linear Review	The process of reviewing documents in sequence, as they naturally appear in a collection. This is the most basic way of reviewing documents.
Batching	The process of gathering or grouping documents for review.
Bates Numbering	Bates numbering is a way of identifying every specific document within an eDisclosure production by assigning a unique, sequential identification number to each page, file, or image.
Big Data	Describes data sets so large and complex that traditional data processing applications/methods cannot be used.
Boolean Search	This logic is used to connect individual keywords or phrases with a single query, used to avoid false positives, and accurately pinpoint documents of interest. Typical connectors are terms such as AND, OR, and NOT.
Chain of Custody / Continuity of Evidence	Continuity of evidence (as it is known in the UK) covers the logical sequence of gathering evidence, whether physical or electronic. Each link (or step) in the chain (or process) is essential. If one of the steps is broken (i.e. a step is missed out), the evidence presented for consideration may be rendered inadmissible. Preserving the continuity of evidence is about following a correct and consistent procedure, and ensuring the quality of evidence brought before the Courts.
Clustering	Clustering assists with understanding general themes and patterns in the early case assessment stage of a disclosure review. Clustering software examines text within documents, determines which documents are related to one another, and groups them into clusters. Clustering allows for the review of document "themes" in the form of visual clusters. Clustering also visually displays patterns of custodian interactions (e.g. by seeing the visual pattern of email traffic).

Coding / Tagging	The method of entering fields of information on a document tagging tree (or coding layout) and saving them in a format that will be linked to that particular document within a document review database. There are different types of coding (which are commonly referred to as 'tagging'): **objective and subjective**. *Objective coding* is used for pre-review categorisation and administration tagging. *Subjective coding* is carried out by document reviewers and for effective coding, and requires an understanding of the matter, together with review instructions.
Computer Assisted Review (CAL)	This method of review utilises advance machine learning, including predictive coding, in order to apply reviewers' coding decisions to large amounts of data. Please see TAR below.
Concept Search	A method of searching for content *not based on keywords*, but on the subject matter of the document, paragraph, or sentence. This is different to keyword searching which requires an exact keyword hit. To provide an example, you might be spoken to about presents, fairy lights, a big turkey, carols, and from that you may gather that the concept being discussed is *Christmas*, even though the word "Christmas" has not been mentioned. In the same way, concept searching trains the technology to recognise associative words to garner the general concept of the document.
Container File	This is a single file that contains multiple other files or documents. A common container file would be a zip file. Container files are typically used due to their considerably smaller file size. Extracted contents are usually anywhere from 50% to 250% larger in size that the original container file. Initial pre-processed data is provided in container files, thus making initial data volumes difficult to exactly gauge.
.csv file	.csv files ("CSV files") are also known as load files. Load files are used to load ESI into a review platform - they are a mandatory requirement alongside a disclosure bundle. To explain further, a load file is similar to an Excel spreadsheet. Information within a load file is used to link a document's metadata to its image or native file. A load file contains the relevant metadata for each document in a disclosure set, i.e. sender, recipient, date sent, subject, etc.
Culling	A broad term that describes the act of removing documents from a collection in an attempt to reduce the size of that collection, and therefore reduce the overall size of the document review population.

	Some standard ways to cull are DeNIST (see below), deduplication ("de-dupe") (see below), applying date ranges, running search terms, and some forms of analytics (such as Technology Assisted Review (TAR), content analysis (checking documents for duplicative content)).
Custodian	The person or entity who "owns" a document. For example, you are the custodian of your emails, and the files on your computer.
Data Extraction	The process of breaking down data from electronic documents to identify their metadata and body contents.
Data Mapping	The process of creating a "map" to identify and record both the location (i.e., recording sophisticated file paths) and the type of information that is available within in an organisation's network and devices.
Deduplication / De-Duping	The process of removing duplicate files from a collection of ESI based on their hash values (see below for a definition of hash values). If two documents, or a family of documents, in a collection have the same hash value, one of them is removed at the processing phase.
DeNIST	The US National Institute of Standards and Technology ("NIST") has a running list of non-user generated document signatures (i.e., computer made files, rather than man-made files). DeNISTED files are industry accepted lists of "junk" files globally and used as a standard in the UK, irrespective of the American origins. When you "DeNIST" a collection of ESI you are simply removing these industry-accepted junk files from the collection and focusing on the man-made content, which will be subject to review and further interrogation.
Document	Any file that contains information. Sometimes a specific file ("Parent") can also have attachments within it ("Children") and that entire group of documents is called a "Family." So, if we have an email that has two attachments, the email is the Parent document, the two attachments are Children documents - together all three are considered a Family. Attachments within attachments are called "Grandchildren"
Early Case Assessment (ECA)	Described by a variety of tools (e.g. early keywords, visual analytics – clustering, etc.) or methods for investigating and quickly learning about the content of a document set.
EDRM	All Disclosure follows the EDRM model - it outlines the stages of the end-to-end Disclosure process. EDRM consists of <u>nine distinct stages</u>, which do not always follow each other in any particular order, i.e., EDRM illustrates an iterative process that can be repeated and completed in a different order. The stages are: information governance; identification; preservation; collection; processing; review; analysis; production; presentation.
E-File	An electronic file, such as an email or word document.
Electronically Stored Information (ESI)	ESI is a document created, altered, communicated, stored, and best utilised in digital form, requiring the use of computer hardware and software to review it.
E-mail Threading	The process of compiling all the emails in a dataset and organising them into conversations. The basic premise is, for example: *Person A* sends *Person B* an email and they reply back and forth to one another a number of times and, in the middle of all that, *Person B* forwards the email to others to join the conversation. Threading can dramatically

	increase review speeds of email data by having the entire conversation reviewed by one reviewer as well as the ability to read the final inclusive email as opposed to all of the conversation pieces separately. Depending on the Disclosure matter at hand, we may ask service providers to sort the emails by the most inclusive email.
Embedded Content	Many file types, including Microsoft Office and Adobe Acrobat files, act as containers (container files) and allow other documents to be linked to them or embedded in them. For example, one can embed a file into an MS Word document by simply dragging it into an open Word document. Depending on the file type and method used, the embedded document may or may not be directly visible in its parent document. For example, the contents of a single-page Visio drawing inserted into an Excel spreadsheet can be visible when the spreadsheet is viewed, while a ZIP file or an MSG file inserted into a Word document would typically be displayed as an icon and its contents would not be directly visible. Extracting embedded objects means that the eDisclosure software identifies each linked or embedded document and extracts it (and its children recursively) as separate records during processing. Additionally, a parent/child relationship is established between the container document and the files embedded in it. We typically ask our eDisclosure providers to extract embedded items as a standard process.
Filtering	The process of using certain criteria to remove documents that do not fit within defined parameters in order to reduce the volume of the overall reviewable data-set.
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. One way to prove that the documents are unaltered is by assigning hash values to the collection of documents, as mentioned below.
Fuzzy Search	Fuzzy search allows searching for word variations such as in the case of misspellings or where language variations or character variations are used. Typically, such searching includes some form of distance and score computations between the specified word and the words in the document review corpus.
Global vs. Custodial Deduplication	There are two different method for conducting deduplication, custodial de-duplication and global de-duplication. Custodial Deduplication removes all duplicate files within a single custodian’s collection. Global Deduplication removes all duplicates across all custodians in a matter (i.e., if *Person A* and *Person B* receive the same email, only one copy of this two-way correspondence will be uploaded to the review platform).

	It is a judgement call as to which preferred method is utilised, depending on the matter at hand. Typically global de-duplication might suit most eDisclosure projects, as it often results in fewer documents to review – on the other hand, custodial deduplication ensures that a custodian’s full collection is kept intact. **We have played out both methodologies in practice by way of example:** In the regular course of business, *Custodian A* emails *Custodian B*, and then *Custodian B* saves a copy of the email to his or her Archive folder. This same email now exists in three places: 1) *Custodian A*’s Outbox; 2) *Custodian B*’s Inbox; and 3) *Custodian B*’s Archive folder. Now, let’s say that both *Custodian A* and *Custodian B* are people of interest in a litigation and both of their emails are collected, but *Custodian A* is a higher priority custodian. **For Custodial Deduplication**: **A copy of the document will exist for each custodian.** If this collection is deduped 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 - but only one copy because the archived copy of the email will be counted as a duplicate and removed. **For Global Deduplication**: **One copy of the document will exist across all custodians.** 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. The reason for the document removal from *Custodian B*'s collection is because *Custodian A* is considered a higher priority custodian, and therefore only one version of a document can exist when globally deduping. *Typically there will be a record kept in the database that Custodian B also had this document.*
Harvesting	Also referred to as the collection of ESI. Harvesting is the method of gathering electronic data for future use whilst maintaining file and system metadata.
Hash Value	A document's digital fingerprint – no two hash values are identical. This is essentially a value that is automatically assigned to data, based on an algorithm. 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. A hash value can be applied to an individual file to identify duplicate files in a collection, or an entire collection of files can be assigned a hash value (to authenticate that the data-set has not been altered by showing that the hash value of the data set has not changed). Typically, but not always, MD5Hash is used as a value for de-duping at the processing stage.

Hosting	Defines a service provided by a technology provider, facilitating access to documents relating to a particular matter within a review software platform.
Image (Forensic Image):	An inclusive, complete and 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, it can be “mounted”, and reviewed in the exact same manner as it would have appeared in its original format pre-collection. One of the biggest advantages of a forensic image is that it captures the “unallocated space” on a computer's drive, which is where deleted fragments of files can still be recovered if they are relevant to an eDisclosure matter.
Keywords	Keywords are exact terms and phrases contained within a dataset, or wildcard or stemmed variations of those terms and phrases. Keywords are run on a review population's extracted text/searchable text. Documents that contain keyword 'hits' can be isolated for priority review. Keywords can be provided in search syntaxes, known as "search strings", and they can use Boolean Logic operators, as well as word proximities, i.e. (("Fish") w/10 ("water")) will return any reference to the word "fish" within ten words of the word "water".
Legacy Data	Data whose format has become obsolete making it difficult to access or process.
Legal Hold / Preservation Order	PF51U, paragraph 4 – Duty to preserve documents.
Load File / .csv file	See CSV file
Metadata	Simply put, metadata is data *about* electronic data. All electronic documents contain other information about a document, which cannot be seen on the surface of the document. In other words, metadata is 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. For example, some basic image files have almost no metadata, but a Microsoft Word document contains hundreds of pieces of metadata, including when the document was created, printed last, and last modified. During processing, as mentioned below, metadata is extracted from a file to make the metadata searchable.
Native	A document that is in its native format, or the format that a document would naturally appear on your computer, such as a Microsoft Word or Excel document. This is opposed to an image format, such as a TIFF format (please see below).
Near-duplicate	Documents that contain a high percentage of the same content are referred to as near-duplicates. During the data reduction process near-duplicates are identified, thus reducing the time and costs associated with review. Near de-duplication, unlike de-duplication, will involve some subjective decisions (i.e., the threshold for near-duplicate similarity. >99% near-textual duplicates may be identified and the 'principal' document reviewed). It is best to discuss, consider and document the implementation of near-duplicate analytics before applying.

Noise Words	Noise words, also known as *stop words*, are words that are not significant indicators for content, e.g. and, who, what, where, when, I, went, etc. Noise words are typically pre-indexed in review databases. Additional noise words may be added to a review database's index (and some may be removed, depending in the situation).
OCR	Optical Character Recognition (“OCR”) is a way to obtain searchable text from a document that does not have any. Essentially, this technology converts a picture or image of text to usable text (extracted text) by scanning over the image to identify a character and recording it as text. Since it works based solely on the appearance of the alphabetic letter or language character, lower quality images can yield bad OCR results. Frequently, similar looking words or letters can get mixed up. For example, the word “Oil” may be recorded as “Oll’ or “Oii”, if the image quality is poor and the OCR software cannot distinguish between an “i” and an “l.”
Personal Storage Table (.pst)	A file format used to host copies of messages, calendar events, and other items within Microsoft software (like Microsoft Outlook, Microsoft Exchange Client, and Windows Messaging) – or, in the most basic of terms, it’s how Outlook stores your email.
Precision and Recall	When using TAR (see below), precision and recall is used to identify how precisely the TAR model is working. Precision and recall work by identifying the number of relevant documents which have been correctly tagged. Documents that are deemed not relevant are not classified during this process. Precision and recall is gauged by either correctly identifying relevant documents or correctly excluding not relevant documents.
Processing	The process by which metadata and searchable text are extracted from a Native file and put into a usable (i.e. searchable) format. Once this data is ingested/uploaded to a database, or review platform, documents may be searched based on the metadata or searchable text, mentioned above. This metadata and text is also what is analysed when analytics are run on documents.
Production	The delivery of documents and ESI, to the opposing side or requesting party, which meets the criteria of the Disclosure request. This typically involves producing the documents on hard drives to the other party(ies) or providing the production via SFTP, together with a load file.
Propagation	Tagging propagation allows reviewers to tag one principal document/family member and have that exact coding automatically replicated to other related documents.
Recall	In search results analysis, recall is the measure of the percent of total number of relevant documents in the quantity returned in the results set (see precision and recall above).
Redact	To redact a document is to deliberately cover portions of the document that are considered privileged, proprietary, or confidential. This is usually done by “blacking-out” or “whiting-out” the text within a document that is to be concealed.
Searchable Text / Extracted Text	The body of a document is made searchable when the extracted text within the document is indexed - the index records where various words are located within a collection of documents. Thereafter, when you run a keyword-based search for a particular term across a body of documents, the search is performed by referencing the index to find which documents contain that search

	term. For that index to be created, the text of the document must first be obtained and recorded. This text can be obtained by extracting it from the document, if the document stores the text, or through OCR (see above).
Search Term Hit Report / STR	A search terms report (STR) simplifies the process of identifying documents that contain a specific group of keywords. Instead of running complicated queries, you can use STRs to enter a list of terms or phrases and then generate a report listing their frequencies in a set of documents. We typically ask the eDisclosure provider to pre-code documents that 'hit' on search terms. The search terms in question can then appear as 'read only' text on the tagging tree.
SFTP	Secure File Transfer Portal
Slip Sheet	A slipsheet, or placeholder, is a cover-sheet that stands in place of a document that has been withheld. A document can be withheld for reasons such as privilege, confidentiality, irrelevance (if the document is part of a family, for which (a) document(s) within that family may be relevant).
Social Discovery	Defined as the discovery of ESI on the various social media sites used today, including but not limited to: Facebook, Twitter, YouTube, LinkedIn, and Instagram.
Spoliation	Defined as the destruction or alteration of data that may be pertinent to a legal matter.
Stemming Specification	Stemming specification is another method for matching word variations. Stemming is the process of finding the root form of a word. The stemming specification will match all morphological inflections of the word, so that if you enter the search term "sing", the stemming matches would include singing, sang, and song.
Structured Data	Data stored in a structured format such as a database. This includes data such as SWIFT transactions, time sheets, etc. Data stored in emails, word documents etc. is known as "unstructured data".
System Files	An electronic file that is part of the operating system. These files are created by the computer, not the user of the computer. The most popular system files on a Windows computer include msdos.sys, io.sys, ntdetect.com and ntldr. These are removed at the De-NISTing stage (see above). These file types should not be provided as part of a production.
Tagging	The process of assigning classifications, such as by relevance or privilege, to one or more documents.
Technology Assisted Review (TAR):	The broad concept of using technology to organise or expedite a document review exercise. The term broadly refers to many methods of technology assistance, including analytics, but it is most frequently associated with Predictive Coding. There are various industry-accepted methods of TAR. Typically, the CAL (continuous active learning) model is used, as it requires 'expert' reviewers to train the review platform to recognise relevant content on an on-going basis; this is particularly helpful where review parameters change during the course of the review.
TIFF/ TIFFing	TIFF (Tagged Image File Format) is simply an image format, like JPEG. “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 memorialises the

	document. TIFFing also can allow for additional features to be added to the document without altering the original version of the document, such as adding bates numbers, confidential designations, and redactions.
Unicode	The code standard that prepares for uniform representation of character sets for all languages. It is also referred to as double-byte language.
Unitisation	The process of splitting image files received in multiple page formats down into individual ‘documents’.
Unstructured Data	Data that is unstructured refers to information that does not exist in the usual row-column database (like structured data does). These text (PDF, MS Word) and multimedia data files, such as webpages, videos, audio files or videos, lack the ability to be organised effectively within a database, hence the name “unstructured”.
Wild Card Search	An asterisk (*) can be used within a word to find variations of a word. So, searching for “risk*” gets you “risk,” “risks,” “risky,” “risked,” “riskier”, and so on.