EDRM Glossary

http://www.edrm.net/resources/glossaries/glossary

Version 1.002, April 22, 2016

The **EDRM Glossary** is EDRM's most comprehensive listing of electronic discovery terms. It includes terms from the specialized glossaries listed below as well as terms not in those glossaries.

The terms are listed in alphabetical order with definitions and attributions where available.

EDRM Collection Standards Glossary

The EDRM Collection Standards Glossary is a glossary of terms defined as part of the EDRM Collection Standards.

EDRM Metrics Glossary

The EDRM Metrics Glossary contains definitions for terms used in connection with the updated EDRM Metrics Model published in June 2013.

EDRM Search Glossary

The EDRM Search Glossary is a list of terms related to searching ESI.

EDRM Search Guide Glossary

The EDRM Search Guide Glossary is part of the EDRM Search Guide. The EDRM Search Guide focuses on the search, retrieval and production of ESI within the larger e-discovery process described in the EDRM Model.

IGRM Glossary

The IGRM Glossary consists of commonly used Information Governance terms.

The Grossman-Cormack Glossary of Technology-Assisted Review

Developed by Maura Grossman of Wachtell, Lipton, Rosen & Katz and Gordon Cormack of the University of Waterloo, the Grossman-Cormack Glossary of Technology-Assisted Review contains definitions for terms used in connect with the discovery processes referred to by various terms including Computer Assisted Review, Technology Assisted Review, and Predictive Coding.

If you would like to submit a new term with definition or a new definition for an existing term, please go to our Submit a Definition page, http://www.edrm.net/23482, and complete and submit the form. We appreciate your contributions!

1

10b(5)

Securities and Exchange Commission regulation governing the rights of shareholders. Many lawsuits by shareholders are filed under Rule 10b(5).

Source: Ibis Consulting, Glossary.

17a4

Securities and Exchange Commission regulation relating to data retention for financial services firms.

Source: Ibis Consulting, Glossary.

A

Ablate

To remove. Used to describe the laser-readable "pits" in the recorded layer of optical disks.

Vinson & Elkins LLP Practice Support, EDD Glossary.

Accept on Zero

A statistical sampling procedure (an acceptance sampling procedure) that draws a random sample of objects from a population and checks each one to determine whether it is a defect. If none of the objects in the sample is found to be defective, then we can conclude with a specifiable level of confidence that there were no more than a specifiable proportion of defects in the original population. Finding zero defects in the sample does not mean that there were zero defects in the population, only that there were no more than a specifiable percentage. One application of this procedure in eDiscovery is to draw a random sample from the population of documents determined by the review to be nonresponsive. The size of the sample is determined by your specified confidence level and by the maximum acceptable percentage of responsive documents that were not retrieved. If none of the documents in the sample is found to be responsive, then we can say with confidence X% that there were no more than Y% responsive documents left behind.

Source: Herb Roitblat, Search 2020: The Glossary.

Source: Herb Roitblat, Predictive Coding Glossary.

Accept on Zero Error

A technique in which the training of a Machine Learning method is gauged by taking a Sample after each training step, and deeming the training process complete when the learning method codes a Sample with 0% Error (i.e., 100% Accuracy).

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Accuracy

The fraction of Documents that are correctly coded by a search or review effort. Note that Accuracy + Error = 100%, and that Accuracy = 100% – Error. While high Accuracy is commonly advanced as evidence of an effective search or review effort, its use can be misleading because it is heavily influenced by Prevalence. Consider, for example, a Document Population containing one million Documents, of which ten thousand (or 1%) are Relevant. A search or review effort that identified 100% of the Documents as Not Relevant and therefore, found none of the Relevant Documents, would have 99% Accuracy, belying the failure of that search or review effort.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Acetate-Base Film

A film substrate used in microfilm production. Considered a safety film (ANSI Standard).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Acrobat

Adobe's electronic document format. Documents can be created from within a word processor, from postscript, or from scanned pages. The documents are highly portable, yet maintain the look of the original. Acrobat is especially useful in this area because Adobe makes the reader available for free. Version 3.0 also makes it integrate well with web browsers.

Source: RSI, Glossary.

See also:

PDF

External link:

Adobe Acrobat Family, http://www.adobe.com/products/acrobat/main.html

Active Data

Data currently displayed on a computer screen, and/or files on a computer that can be accessed without having to use a restoration process.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Richard A. Lazar, The Guide to Electronic Discovery, at 37 (Fios, Inc. 2002).

The information readily available and accessible to users, including word processing files, spreadsheets, databases' data, e-mail messages, electronic calendars and contact managers.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Feldman, The Essentials of Computer Discovery, Computer Forensics Inc. (1/1/2001), http://www.forensics.com/pdf/Essentials_of_Discovery.pdf#page=2.

Active data is information residing on the direct access storage media of computer systems, which is readily visible to the operating system and/or application software with which it was created and immediately accessible to users without undeletion, modification or reconstruction (i.e., word processing and spreadsheet files, programs and files used by the computer's operating system).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms.

Active data is information residing on the direct access storage media of computer systems, which is readily visible to the operating system and/or application software with which it was created and immediately accessible to users without undeletion, modification or reconstruction.

Source: Merrill Corporation, Electronic Discovery Glossary.

Data existing on the data and file storage media of computer systems. Active data is easily viewed on the operating system and/or application software that was used to create it and is directly available to users without un-deletion, alteration, or restoration.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Data currently displayed on a computer screen.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html.

Source: RSI, Glossary.

Information residing on the computer which is visible and fully available to the user.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Active Learning

An Iterative Training regimen in which the Training Set is repeatedly augmented by additional Documents chosen by the Machine Learning Algorithm, and coded by one or more Subject Matter Expert(s).

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A form of supervised machine learning that presents for review or human categorization the documents with the highest current uncertainty, those documents that will be most informative about how to update the learning process.

Source: Herb Roitblat, Search 2020: The Glossary.

Source: Herb Roitblat, Predictive Coding Glossary.

Active Record

Active records are records related to current, ongoing or in process activities and are referred to on a regular basis to respond to day-to-day operational requirements. An active record resides in native application format and is accessible for purposes of business processing with no restrictions on alteration beyond normal business rules.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms.

Activity

Single logical query or the progression of single logical queries performed interactively in an effort to accumulate intelligence.

Source: EDRM Search Glossary.

Ad Hoc Search

Single logical query or the progression of single logical queries performed interactively in an effort to accumulate intelligence.

Source: EDRM Search Glossary.

See also:

Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Ad Hoc Workflow

A simple manual process by which documents can be moved around a multi-user imaging system on an "as-needed" basis.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Ad Hoc Workflow

Rule-Based Workflow

Workflow

Adaptive Pattern Recognition

The system indexes every letter on every page. When the user conducts a search, the system conducts a search based on discrete patterns in the text.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search

Adaptive pattern recognition

Associative retrieval

Boolean search

Combined word search

Compliance Search

Concept search

Exploratory Search

Full text search

Fuzzy search

Index

Index/coding field

Keyword

Keyword search

Natural language search

Numeric range search

Phonic search

Phrase search

Proximity search

Range search

Search

Similar document search

Sound-alike

Stemming

Synonym search

Term search

Topical search

Weighted relevance search

Wildcard search

ADC (Analog to Digital Converter)

Changes analog signals to digital representations (numbers).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Additive Color

All the colors in the light spectrum add up to make white light. Computer monitors use a three additive colors, Red, Green & Blue (RGB).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

ADF (Automatic Document Feeder)

A device that holds pages and feeds them one after another into a scanner.

Source: RSI, Glossary.

This is the means by which a scanner feeds the paper document.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Admissible

Evidence that is acceptable or allowable in court.

Source: RenewData, Glossary (10/5/2005).

Admissible evidence in a court of law is any testimonial, documentary, or tangible evidence that may be introduced to a fact finder--usually a judge or jury--in order to establish or to bolster a point put forth by a party to the proceeding. In order for evidence to be admissible, it must be relevant, without being prejudicial, and it must have some indicia of reliability.

For evidence to be relevant, it must tend to prove or disprove some fact that is at issue in the proceeding. However, such evidence will not be admissible if the utility of the evidence is outweighed by its tendency to cause the fact finder to disapprove of the party it is introduced against for some unrelated reason. Furthermore, certain public-policy considerations bar the admission of otherwise relevant evidence.

For evidence to be reliable enough to be admitted, the party proffering the evidence must be able to show that the source of the evidence makes it so. If the evidence is in the form of witness testimony, the party introducing the evidence must lay the groundwork for the credibility of the witness, and his knowledge of the things to which he attests. Hearsay is generally barred for its lack of reliability. If the evidence is documentary, the party proffering the evidence must be able to show that it is authentic and must be able to demonstrate the chain of custody from the original author to the present holder.

The trial judge performs a "gatekeeping" role in excluding unreliable testimony. The United States Supreme Court first addressed the reliability requirement for experts in the landmark case *Daubert v. Merrell Dow Pharmaceuticals, Inc.* 509 U.S. 579 (1993). The Court laid out four non-exclusive factors that trial courts may consider when evaluating scientific expert reliability: (1) whether scientific evidence has been tested and the methodology with which it has been tested; (2) whether the evidence has been subjected to peer review or publication; (3) whether a potential rate of error is known; and (4) whether the evidence is generally accepted in the

scientific community. *Id*. at 592-94. *Kumho Tire Co., Ltd. v. Carmichael* later extended the *Daubert* analysis to include all expert testimony. 526 U.S. 137 (1999).

Source: EDRM Presentation Guide.

Agreement

The fraction of all Documents that two reviewers code the same way. While high Agreement is commonly advanced as evidence of an effective review effort, its use can be misleading, for the same reason that the use of Accuracy can be misleading. When the vast majority of Documents in a Population are Not Relevant, a high level of Agreement will be achieved when the reviewers agree that these Documents are Not Relevant, irrespective of whether or not they agree that any of the Relevant Documents are Relevant.

Source: The Grossman-Cormack Glossary of Technology Assisted Review (Version 1.02, Nov. 2102).

AI (Artificial Intelligence)

See: Artificial Intelligence

AIIM (Association for Information and Image Management)

The Association for Information and Image Management – focused on electronic imaging.

External links:

AIIM, http://www.aiim.org/

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Algorithm

A mathematical set of steps designed to solve a problem or run instructions in a program. For example, an algorithm in a case management program would perform the function, “Check the filing date of the complaint in this matter, determine the date to file an answer, determine if the answer has been sent out and, if not, send email to the attorney in charge of the case warning him of the impending date.”

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A formally specified series of computations that, when executed, accomplishes a particular goal. The Algorithms used in E-Discovery are implemented as computer software.

Source: The Grossman-Cormack Glossary of Technology Assisted Review (Version 1.02, Nov. 2102).

A specific set of steps that when accurately executed leads to a specific outcome. Algorithms can be created for many different kinds of processes, including calculation, data processing, automated reasoning, and mathematical computations. Algorithms should be distinguished

from heuristics. The word “algorithm” is often misused to refer to any computer-implemented process.

Source: Herb Roitblat, Search 2020: The Glossary.

Aliasing

When computer graphics output has jagged edges or a stair stepped appearance when magnified. Homonym is "anti-aliasing”.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Alpha

The complement of confidence level. 1 – confidence level. Statisticians refer to the alpha level to decide whether a value is significant. A 95% confidence level has a 5% alpha level. A 99% confidence level has a 1% alpha level.

Source: Herb Roitblat, Predictive Coding Glossary.

Alphanumeric

Set of characters composed of letters and numbers; may include punctuation marks or other symbols; excludes printer control characters such as "carriage return" and flow control characters such as XON and XOFF.

Source: RSI, Glossary.

Characters composed of letters, numbers (and sometimes punctuation marks). Excludes printer/flow control characters, (Carriage Return/XON & XOFF).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

ALS (Automated Litigation Support)

The process of using computers to control data during litigation.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Ambient Data

Data stored in non-traditional computer storage areas and formats, such as Windows swap files, unallocated space and file slack.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

See also:

Ambient data	Free space	Slack space
Fragmented data	Residual data	Swap file

Unallocated space

American National Standards Institute (ANSI)

The American National Standards Institute, or ANSI, "is a private, non-profit organization (501(c)3) that administers and coordinates the U.S. voluntary standardization and conformity assessment system. The Institute's mission is to enhance both the global competitiveness of U.S. business and the U.S. quality of life by promoting and facilitating voluntary consensus standards and conformity assessment systems, and safeguarding their integrity."

Source: http://www.ansi.org/about_ansi/overview/overview.aspx?menuid=1

External link:

American National Standards Institute, http://www.ansi.org

American Standard Code for Information Interchange (ASCII)

Allocates a number to each key on the keyboard that can be traded and read by most computer systems. A text file.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

ASCII: The acronym for the American Standard Code for Information Interchange, which has assigned a coded set of numbers to represent letters and other special characters. ASCII data consists only of text with no formatting (e.g. bold or italics).

Source: Ibis Consulting, Glossary.

A standard code used for data exchange between computers. An ASCII (pronounced “as-key “) text file contains only the letters of the alphabet, numbers, punctuation, and certain communications symbols, but no embedded word-processing codes. An ASCII data file (or ASCII delimited file) has the data in fields that are separated by quotation marks or commas and that allows easy transfer into a database or spreadsheet.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Pronounced ask-ee. American Standards Committee II. An eight bit computer coding structure for letters, numbers and characters in which seven bits are used to identify each individual entity (128 maximum), with one bit for parity. When no parity bit is used, all eight bits can be used to represent up to 256 characters; this character set is extended ASCII.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

ASCII is a code that assigns a number to each key on the keyboard. ASCII text does not include special formatting features and therefore can be exchanged and read by most computer systems.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The code by which English letters are represented inside a computer. Most commonly used to discuss a document from which all formatting information (other than spaces and paragraph breaks) has been removed. The text of those documents. MS Word documents, for example, include a lot of information in addition to the text that specifies how the document should look, revisions, and so forth. The so-called ASCII text of this document just contains the text of the document, with everything else removed.

Analog

The electrical replica or waveform of a physical process caused by changes in amplitude or frequency. Opposite of digital (Zeros & Ones).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Analog to Digital Converter

See: ADC (Analog to Digital Converter)

Analysis Phase

Evaluating ESI for content and context, including key patterns, topics, people and discussion.

Source: EDRM Stages

Corresponds to UTBMS Code L660. Activities and actions required by litigation teams to be able to make informed decisions about strategy and scope through reliable methods based on verified data.

Source: EDRM Metrics Glossary

Annotation

A note placed in a full-text record to comment on the textual material.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The changes or additions made to a document using sticky notes, a highlighter, or other electronic tools. Document images or text can be highlighted in different colors, redacted (blacked-out or whited-out), stamped (e.g. “FAXED” or “CONFIDENTIAL”), or have electronic sticky notes attached. Annotations should be overlaid and not change the original document.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

ANSI

See: American National Standards Institute (ANSI)

Aperture Card

An IBM punch card with a window which holds a 35mm frame of microfilm. Indexing information is punched in the card.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Applet

An application program that uses the client's web browser to provide a user interface.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Application

A program, that instructs a computer to perform a specific set of instructions or execute a process. Some software applications are user-driven like Microsoft Word or Notepad, while others are system-driven like the Windows system clock or automatic virus scanning programs.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

An application is a collection of one or more related software programs that enables a user to enter, store, view, modify or extract information from files or databases. The term is commonly used in place of “program,” or “software.” Applications may include word processors, Internet browsing tools and spreadsheets.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A set of electronic instructions, also known as a program, which instructs a computer to perform a specific set of processes.

Source: RSI, Glossary.

A program that performs “people” functions, such as word processing, spreadsheets, or litigation support.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A program used to generate documents (e.g., MS Word, Word Perfect, Pagemaker, Visio). Each application is typically associated with one or more file extensions (e.g., .doc, .xls).)

Application File

Computer files that run software applications (such as MS Office, Lotus WordPro and Lotus 1-2-3, Adobe Acrobat, TXT files, TIFFs, etc.) not associated with mail containers or their messages, attachments, or non-mail items.

Source: Ibis Consulting, Glossary.

Application File User

The user data set containing application files.

Source: Ibis Consulting, Glossary.

Application Service Provider (ASP)

An application service provider is a company that delivers software applications to multiple users over the Internet or other network. Instead of purchasing software licenses directly from vendors or re-sellers, companies rent the software from an ASP, which hosts, maintains and upgrades software applications and computer hardware.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Providing a computer program or application across a broadband connection as a third-party provider. Allows users to lower the cost of deploying an application.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Application Software

See: Application

Architecture

The design or physical structure of the computer's internal components and how they work.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Archival Data

Archival data is information that is not directly accessible to the user of a computer system but that the organization maintains for long-term storage and record keeping purposes. Archival data may be written to removable media such as a CD, magneto-optical media, tape or other electronic storage device, or may be maintained on system hard drives in compressed formats (i.e., data stored on backup tapes or disks, usually for disaster recovery purposes).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Information that is not directly accessible to the user of a computer system but that the organization maintains for long-term storage and record-keeping purposes. Archival data may be written to removable media such as a CD, magneto-optical media, tape or other electronic storage device, or may be maintained on system hard drives in compressed formats.

Source: Merrill Corporation, Electronic Discovery Glossary.

Data that is not immediately available to the computer user but that the organization preserves for storage and record keeping purposes, often stored on CD-ROMs, tapes, or other electronic storage devices.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Information that is not directly available to the user of a computer but has been stored on the computer system and can be retrieved through a special process.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Archive

A long-term computer storage area.

Source: RenewData, Glossary (10/5/2005).

Archives are long term repositories for the storage of records. Electronic archives preserve the content, prevent or track alterations and control access to electronic records.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A copy of data on a computer drive, or on a portion of a drive, maintained for historical reference.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Fios's eDiscovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

A container that holds files, either compressed or uncompressed (ZIP, CAB, TAR, GZ, JAR, PST, NSF, or other file types). There are two types of archives – mail containers and file containers.

Source: Ibis Consulting, Glossary.

The procedure of transferring text or data from a hard disk to off-line storage media for later access.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Area Under the ROC Curve (AUC)

From Signal Detection Theory, a summary measure used to assess the quality of Prioritization. AUC is the Probability that a randomly chosen Relevant Document is given a higher priority than a randomly chosen Non-Relevant Document. An AUC score of 100% indicates a perfect ranking, in which all Relevant Documents have higher priority than all Non-Relevant Documents. An AUC score of 50% means the Prioritization is no better than chance.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Artificial Intelligence (AI)

A category of computer science dealing with the ability of machines to perform in a manner associated with human beings, such as reasoning, learning, or understanding language. Currently associated with voice recognition technology and, to a lesser degree, optical character recognition (OCR).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

An umbrella term for computer methods that emulate human judgment. These include Machine Learning and Knowledge Engineering, as well as Pattern Matching (e.g., voice, face, and handwriting recognition), robotics, and game playing.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

ASCII

See: American Standard Code for Information Interchange (ASCII)

ASP

See: Application Service Provider (ASP)

Aspect Ratio

The relationship of the height and width of any image. This must always be preserved to prevent distortion.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Aspects

The major elements common to each e-discovery Phase around which identifiable metrics activity aggregates: Custodians, Systems, Media, Status, Format, QA & Control, and Activities.

Source: EDRM Metrics Glossary

ASR

See: Automated Speech Recognition (ASR)

Asset

The specific container of information that IT stores and secures under their management. The primary driver is to increase "Efficiency" and lower costs associated with this function.

Source: IGRM White Paper

Association for Information and Image Management, The

See: AIIM (Association for Information and Image Management)

Associative Retrieval

When certain terms appear frequently in the vicinity of the terms for which the user is searching, these associative words may provide clues for further searching.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Attachment

Any file type associated with or attached to an e-mail.

Source: RenewData, Glossary (10/5/2005).

A memorandum, letter, spreadsheet, or any other electronic document appended to another document or email.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

Files attached to mail message (or sometimes embedded into mail message).

Source: Ibis Consulting, Glossary.

An enclosure to a transmittal letter or an exhibit to a primary document.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Any electronic document appended to another document, typically email.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

An attachment is a record or file associated with another record for the purpose of storage or transfer. There may be multiple attachments associated with a single "parent" or "master" record. The attachments and associated record may be managed and processed as a single unit. In common use, this term refers to a file (or files) associated with an e-mail for transfer and storage as a single message unit. Because in certain circumstances the context of the attachment—for example, the parent e-mail and its associated metadata—can be important, an organization should consider whether its policy should authorize or restrict the disassociation of attachments from their parent records.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Attachment Field

A data field used to record information about enclosures and/or attachments to a "parent" document.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Attorney Notes Field

A data field used for ongoing attorney notes and comments.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Attribute

A data attribute is a characteristic of data that sets it apart from other data, such as location, length, or type. The term attribute is sometimes used synonymously with “data element” or “property.”

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Audio-Video Interleave (AVI)

A Microsoft standard for Windows animation files. The format interleaves audio and animation to provide medium quality multimedia.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Audit Trail

In computer security systems, a chronological record of when users logged in, how long they were engaged in various activities, what they were doing, and whether any actual or attempted security violations occurred. An audit trail is an automated or manual set of chronological records of system activities that may enable the reconstruction and examination of a sequence of events and/or changes in an event.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Authentication

Authentication is the act of establishing or confirming something (or someone) as authentic. This might involve confirming the identity of a person, the origins of an artifact, or assuring that a computer program is a trusted one.

Source: EDRM Presentation Guide.

Author

The author of a document is the person, office or designated position responsible for its creation or issuance. In the case of a document in the form of a letter, the author or originator is usually indicated on the letterhead or by signature. In some cases, the software application producing the document may capture the author's identity and associate it with the document.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Author Field

A data field used for recording names of individuals and/or business entities who wrote, sent, or transmitted a document.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Auto-Categorization

The process of using machine learning or other rule-based systems for categorizing documents without direct human intervention. For example, emails may be auto-categorized as they arrive at an archive as to their retention period. The categories may be based on a taxonomy or ontology.

Source: Herb Roitblat, Search 2020: The Glossary.

Auto Function

This function dynamically updates field code values entered by application users in Microsoft Office and AutoCAD documents.

Source: Ibis Consulting, Glossary.

Autoexec.bat

Usually pronounced autoexecdotbat, this is a special batch file used on PCs that runs when the computer is turned on and tells the computer what programs to execute first.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Automated Litigation Support (ALS)

See: ALS (Automated Litigation Support)

Automated Speech Recognition (ASR)

Also called automated voice recognition (AVR). A program that will "translate" words spoken into a microphone connected to a computer into written text in a word-processing program or perform a function in a database program.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Automated Voice Recognition (AVR)

See: Automated Speech Recognition (ASR)

Automatic Document Feeder (ADF)

See: ADF (Automatic Document Feeder)

AVI

See: Audio-Video Interleave (AVI)

AVR

See: Automated Speech Recognition (ASR)

B

Back-End / Front-End

Expressions that describe programs relative to the user. A front-end program is one that users interact with directly, while a back-end program supports the front-end services.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Backfile

An existing paper or microfilm file.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Backup

To create a copy of data as a precaution against the loss or damage of the original data. Most users backup some of their files, and many computer networks utilize automatic backup software to make regular copies of some or all of the data on the network. Some backup systems use digital audio tape (DAT) as a storage medium.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A duplicate of information as a preventative measure against the potential loss of data that is done regularly by many computer users. Many organizations also utilize automatic backup software that regularly stores data.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A copy of inactive data, intended for use in the restoration of data lost to catastrophic failure of system memory.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

See also:

Backup
Backup tape
DAT - digital audio tape
Data extraction
Digital audio tape
Disaster recovery tape
DLT - digital linear tape
Magnetic storage media
Media
QIC - quarter inch cartridge
Tape

Backup Data

Backup data is information that is not presently in use by an organization and is routinely stored separately upon portable media, to free up space and permit data recovery in the event of disaster.

Source: Merrill Corporation, Electronic Discovery Glossary.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Information stored separately from the computer system to permit data recovery in the event of disaster.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Backup Tape

Tape devices that transfer active data to inactive data, intended for use in data restoration. Backup tapes typically use data compression, which makes restoration time-consuming and expensive, especially given the lack of uniform standards governing data compression.

Source: Ibis Consulting, Glossary.

Backup or disaster recovery tapes are portable media used to store data that is not presently in use by an organization to free up space but still allow for disaster recovery.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Tape media used to back up data.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

One of various types of magnetic recording tapes that are used to save a snapshot of the current state of a file system. Backup tapes are designed to be able to restore the content of a file system if it should become corrupted.

See also:

Backup
Backup tape
DAT - digital audio tape
Data extraction
Digital audio tape
Disaster recovery tape
DLT - digital linear tape
Magnetic storage media
Media
QIC - quarter inch cartridge
Tape

Backup Tape Recycling

The process whereby an organization's backup tapes are overwritten with new archived data usually on a fixed schedule (e.g., the use of nightly backup tapes for each day of the week with the daily backup tape for a particular day being overwritten on the same day the following week; weekly and monthly backups being stored offsite for a specified period of time before being placed back in the rotation).

Source: Merrill Corporation, Electronic Discovery Glossary.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Applied Discovery's Bone up on Backup, http://www.lexisnexis.com/applieddiscovery/NewsEvents/PDFs/BoneUpOnBackup.pdf

Backup tape recycling is the process whereby an organization’s backup tapes are overwritten with new backup data, usually on a fixed schedule (i.e., the use of nightly backup tapes for each day of the week with the daily backup tape for a particular day being overwritten on the same

day the following week; weekly and monthly backups being stored offsite for a specified period of time before being placed back in the rotation).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The act by which old backup tapes are overwritten with new data.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary

Bag and Tag

The process of receiving, recording, and securing client source data as evidence. The first link in the chain of custody.

Source: Ibis Consulting, Glossary.

Bag of Words

A Feature Engineering method in which the Features of each Document comprise the set of words contained in that Document. Documents are determined to be Relevant or Not Relevant depending on what words they contain. Elementary Keyword Search and Boolean Search methods, as well as some Machine Learning methods, use the Bag of Words model.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Bandwidth

The quantity of information that can be sent over a network in a certain amount of time.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The amount of information or data that can be sent over a network connection in a given period of time. Bandwidth is usually stated in bits per second (bps), kilobits per second (kbps), or megabits per second (mps).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Bar Code

A small pattern of vertical lines that is read by a laser or an optical scanner, and which corresponds to a record in a database. An add-on component to imaging software, this feature is designed to increase the speed with which documents can be archived.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Basic De-Duplication

Performed on a select and limited basis, such as for file names and types, and is usually based on the hash value of the entire electronic document.

Source: RenewData, Glossary (10/5/2005).

See also:

Case de-duplication	Duplicate	Horizontal Deduplication
Custodian de-duplication	Dynamic de-duplication	Production de-duplication
De-duplication	Global Deduplication	Vertical Deduplication

Basic Input Output System (BIOS)

A special program contained in the computer's ROM that controls the components of the computer and how they interact and work together.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The specific PC input/output "rules" and the programs which execute these to allow the transfer of information to/from the "central processing unit" of the PC.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Batch

A file containing one or more commands that execute consecutively, one at a time.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A collection of material for input into the computer, such as a batch of documents segregated for coding or a batch of data records to be restored from a backup tape.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Batch Printing

The process of printing a group of documents, usually from the TIFF's or PDF's.

Batch Processing

The name of the technique used to input a large amount of information in a single step, as opposed to individual processes.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Batchload

A client-specific document load file, generating an output directory structure for deliverables (usually TIFF or PDF images, metadata and text representations of files, but sometimes files in their native format).

Source: Ibis Consulting, Glossary.

Bates Number

A document identification technique in which every page (or image) of every document in a document collection is assigned a unique, sequential identification number. Bates numbers may be then printed onto the document page before the page is distributed to multiple parties to ensure that each distributed page can be identified and compared to the original.

Source: Ibis Consulting, Glossary.

The Bate® number is a number that uniquely identifies each page of a document.

Source: RSI, Glossary.

A bates production number is a tracking number assigned to each page of each document in the production set.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A unique number that is attached to each page of a document (in electronic or manual form) to identify it. The word Bates comes from the Bates Company, which was one of the originators of numeric (and alpha) stamping machines.

See also:

Bates number
Bates prefix
Bates stamp
Bates stamping
Document number

Bates Prefix

A project-specific, client specification in the form of an alphanumeric prefix that precedes a project's control number (the digital equivalent of a Bates number). Also called "control number prefix."

Source: Ibis Consulting, Glossary.

See also:

Bates number
Bates prefix
Bates stamp
Bates stamping
Document number

Bates Stamp

See also:

Bates number
Bates prefix
Bates stamping
Document number

Bates Stamping

The process of adding an image representation of a Bates number to a page or image (this includes PDFs and TIFFs).

Source: Ibis Consulting, Glossary.

See also:

Bates number	Bates stamp	Document number
Bates prefix	Bates stamping	

Baud

A unit of data-transmission speed used in discussing modems. One baud equals one bit per second (bps). Divide by 10 to get characters per second (e.g., a 9600 baud modem sends data at 960 characters per second).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Baud	Microcomputer	Personal computer
File server	Minicomputer	Workstation
Laptop computer	Notebook computer	

Baud Rate

See also:

Baud	Microcomputer	Personal computer
File server	Minicomputer	Workstation
Laptop computer	Notebook computer	

Bayes / Bayesian / Bayes' Theorem

A general term used to describe Algorithms and other methods that estimate the overall Probability of some eventuality (e.g., that a Document is Relevant), based on the combination of evidence gleaned from separate observations. In Electronic Discovery, the most common evidence that is combined is the occurrence of particular words in a Document. For example, a Bayesian Algorithm might combine the evidence gleaned from the fact that a Document contains the words "credit," "default," and "swap" to indicate that there is a 99% Probability that the Document concerns financial derivatives, but only a 40% Probability if the words "credit" and "default," but not "swap," are present. The most elementary Bayesian Algorithm is Naïve Bayes; however most Algorithms dubbed "Bayesian" are more complex. Bayesian Algorithms are named after Bayes' Theorem, coined by the 18th century mathematician,

Thomas Bayes. Bayes' Theorem derives the Probability of an outcome, given the evidence, from: (i) the probability of the outcome, independent of the evidence; (ii) the probability of the evidence, given the outcome; and (iii) the probability of the evidence, independent of the outcome.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Bayesian Categorizer

An information retrieval tool that computes the probability that a document is a member of a category from the probability that each word is indicative of each category. These estimates are derived from example documents. Uses the probability of each word given each category to compute the probability of each category given each word. Also called a naïve Bayesian Categorizer.

Source: Herb Roitblat, Predictive Coding Glossary.

Bayesian Classifier

Bayesian classifier is a process of identifying concepts using a certain representative documents in a particular category. The classifier has the ability to discern other responsive documents in the larger collection and place them in a category. Typically, a category is represented by a collection of words and their frequency of occurrence within the document. The probability that a document belongs to a category is based on the product of each word of the document appearing in that category across all documents. Thus, the learning classifier is able to apply words present in a sample category and apply that knowledge to other new documents. In the e-discovery context, a Bayesian classifier can quickly place documents into confidential, privileged, responsive documents and other well-known categories.

Source: EDRM Search Glossary

Bayesian Classifier / Bayesian Filter / Bayesian Learning

A colloquial term used to describe a Machine Learning Algorithm that uses a Bayesian Algorithm Naïve Bayes.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

BBS (Bulletin Board System)

A bulletin board system (BBS) is a computer or an application dedicated to the sharing or exchange of messages or other files on a network. Originally an electronic version of the type of bulletin board found on the wall in many kitchens and work places, the BBS was used to post simple messages between users. The BBS became the primary kind of online community through the 1980s and early 1990s, before the World Wide Web arrived.

Source: WhatIs.com definition, bulletin board system (BBS), http://whatis.techtarget.com/definition/bulletin-board-system-BBS

A bulletin board system, or BBS, is a computer server running custom software that allows users to connect to the system using a terminal program. Once logged in, the user can perform functions such as uploading and downloading software and data, reading news and bulletins, and exchanging messages with other users through email, public message boards, and sometimes via direct chatting. Many BBSes also offer on-line games, in which users can compete with each other, and BBSes with multiple phone lines often provide chat rooms, allowing users to interact with each other. Bulletin board systems were in many ways a precursor to the modern form of the World Wide Web, social networks and other aspects of the Internet. Low-cost, high-performance modems drove the use of online services and BBSes through the early 1990s. Infoworld estimated there were 60,000 BBSes serving 17 million users in the United States alone in 1994, a collective market much larger than major online services like CompuServe.

Source: Wikipedia, Bulletin board system, https://en.wikipedia.org/wiki/Bulletin_board_system

Beginning Document Number

The first page of a document or record. Often BegDoc#.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Beginning Number Field

A data field for recording the number of the first page of a document. Also used as a document identifier to find hard copies or to retrieve images.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Bibliographic / Objective Coding

Objective information, often manually recorded from documents such as the document date, the authors or recipients of the documents, or the title of a document. Bibliographic coding usually takes place against documents originating as paper with no electronically stored information.

Source: EDRM Search Glossary.

The entering of objective information such as date, document number, and document type into data fields.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Extracting information from electronic documents such as date created, author recipient, CC and linking each image to the information in pre-defined objective fields. In direct opposition to Subjective coding where legal interpretations of data in a document are linked to individual documents. Also called objective coding.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary

See also:

Bibliographic Coding
Coding
Indexing
Issue Code
Issue coding
Level coding
Objective coding
Subjective coding
Tag
Taxonomic coding
Verbatim coding

Big Data

An ill-defined term for large collections of data of various sorts. Big data may be big because it includes a large number of records (e.g., all of the transactions on Amazon), because it includes a large number of variables (all of the characteristics or features that a bank knows about each customer), or both. Big data often suffers from the 4-Vs: Velocity, Variety, Volume, and Veracity. Big data accumulate very rapidly, they consist of many different kinds of data, at high volume, and the quality of the data often presents a challenge. Big data can also refer to very large collections of electronically stored information.

Source: Herb Roitblat, Search 2020: The Glossary.

Bigram

An N-Gram where N = 2 (i.e., a 2-gram).

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Binary

Mathematical base 2, or numbers composed of a series of zeros and ones. Since zero's and one's can be easily represented by two voltage levels on an electronic device, the binary number system is widely used in digital computing.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Binomial Calculator / Binomial Estimation

A statistical method used to calculate Confidence Intervals, based on the Binomial Distribution, that models the random selection of Documents from a large Population. Binomial Estimation is generally more accurate, but less well known, than Gaussian Estimate. A Binomial Estimate is substantially better than a Gaussian Estimate (which, in contrast, relies on the Gaussian or Normal Distribution) when there are few (or no) Relevant Documents in the Sample. When there are many Relevant and many Non-Relevant documents in the Sample, Binomial and Gaussian Estimates are nearly identical.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Binomial Distribution

The Probability that a Random Sample from a large Population will contain any particular number of Relevant Documents, given the Prevalence of Relevant Documents in the Population. Used as the basis for Binomial Estimation.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Binomial Estimate

A Statistical Estimate of a Population characteristic using Binomial Estimation. It is generally expressed as a Point Estimate accompanied by a Margin of Error and a Confidence Level, or as a Confidence Interval accompanied by a Confidence Level.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

BIOS

See: Basic Input Output System (BIOS)

Bit

A bit is the smallest unit of information recognized by a computer; it corresponds to a choice between one and zero, the basis for all information storage in binary language computers. Eight bits make up a byte.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A measurement of data. It is the smallest unit of data. A bit is either the "1" or "0" component of the binary code. A collection of bits is put together to form a byte.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterm

Single position in base 2 arithmetic (2 n) – either on (1) or off (0).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Byte	GB - gigabyte	EB - exabyte
KB - kilobyte	TB - terabyte	
MB - megabyte	PB - petabyte	

Bit Map

Bitmap images, also called raster or paint images, are made of individual dots called pixels (picture elements) that are arranged and colored differently to form a pattern. When you zoom in, you can see the individual squares that make up the total image. Increasing the size of a bitmap has the effect of increasing individual pixels, making lines and shapes appear jagged. Reducing the size distorts the original image because pixels are removed to reduce the overall

image size. Because a bitmap is created as a collection of arranged pixels, its parts cannot be manipulated (e.g., moved) individually.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Creating characters or images by creating a "picture" (matrix) of individual bits (pixels). The individual bits may just be binary (black and white) or high definition color. In color systems, the "z-axis" of each pixel has a value which represents the "shade of gray" or color of the bit. This value can be as high as 32 bits for very high resolution color. This results in a large, uncompressed file. For instance, a 300 dpi, E-Size drawing bit map is approximately 16MB.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Bit-by-Bit Copy

See: Bitstream Copy

Bitmap

Bitmap images, also called raster or paint images, are made of individual dots called pixels (picture elements) that are arranged and colored differently to form a pattern. When you zoom in, you can see the individual squares that make up the total image. Increasing the size of a bitmap has the effect of increasing individual pixels, making lines and shapes appear jagged. Reducing the size distorts the original image because pixels are removed to reduce the overall image size. Because a bitmap is created as a collection of arranged pixels, its parts cannot be manipulated (e.g., moved) individually.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Creating characters or images by creating a "picture" (matrix) of individual bits (pixels). The individual bits may just be binary (black and white) or high definition color. In color systems, the "z-axis" of each pixel has a value which represents the "shade of gray" or color of the bit. This value can be as high as 32 bits for very high resolution color. This results in a large, uncompressed file. For instance, a 300 dpi, E-Size drawing bit map is approximately 16MB.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Bitonal

An image or file comprised of pixel or dot values of either black or white.

Source: RSI, Glossary.

Bi-tonal (black and white only, one bit per pixel). A Bi-tonal image is created by a thresholding process from a grayscale input, either during the scanning process or subsequently. Thresholding is an irreversible process which results in speckled images with noticeably "stair-stepped" diagonal lines.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Bits Per Inch (BPI)

This defines data densities in disk and magnetic tape systems.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Bits Per Second (BPS)

In data communications, bits per second (abbreviated bps or bit/sec) is a common measure of data speed for computer modems and transmission carriers. As the term implies, the speed in bps is equal to the number of bits transmitted or received each second.

Source: TechTarget definition, bits per second (bps or bit/sec), http://searchnetworking.techtarget.com/definition/bits-per-second

Bitstream Copy

Bit stream backup (also referred to as mirror image backup) involves the backup of all areas of a computer hard disk drive or another type of storage media. Such a backup exactly replicates all sectors on a given storage device. Thus, all files and ambient data storage areas are copied. Bit stream backups - sometimes also referred to as "evidence grade" backups - differ substantially from traditional computer file backups and network server backups.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing NTI's Computer Forensics Definitions, http://www.forensics-intl.com/def2.html

Bitstream Image

A sector-by-sector, bit-by-bit copy of a physical hard drive or a logical drive.

Source: EDRM Collection Standards

See Bitstream copy: Bit stream backup (also referred to as mirror image backup) involves the backup of all areas of a computer hard disk drive or another type of storage media. Such a backup exactly replicates all sectors on a given storage device. Thus, all files and ambient data storage areas are copied. Bit stream backups - sometimes also referred to as "evidence grade" backups - differ substantially from traditional computer file backups and network server backups.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing NTI's Computer Forensics Definitions, http://www.forensics-intl.com/def2.html

Blair and Maron

Authors of an influential 1985 study study (David C. Blair & M.E. Maron, An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System, 28 COMMC'NS ACM 289 (1985)), showing that attorneys supervising skilled paralegals believed they had found at least

75% of the Relevant Documents from a Document Collection, using Search Terms and iterative search, when they had in fact found only 20%. That is, the searchers believed they had achieved 75% Recall, but had achieved only 20% Recall. In the Blair and Maron study, the attorneys and paralegals used an iterative approach, examining the retrieved Documents and refining their search terms until they believed they were done. Many current commentators incorrectly distinguish the Blair and Maron study from current iterative approaches, failing to note that the Blair and Maron searchers did in fact refine their search terms based on their review of the Documents that were returned in response to their queries.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Blog

A web log. A journal available on a web page, typically on a specific subject and updated daily. Legal blogs are sometimes called “Blawgs.”

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Blogs, also referred to as Web logs, are frequent, chronological Web publications consisting of links and postings. The most recent posting appears at the top of the page.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Blowback

Printing electronic files to paper for review or production in hardcopy form.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Albert Barsocchini, Data Collection Standards (LTN 1/15/04), http://www.law.com/special/supplement/e_discovery/data_collection.shtml

The to-be-printed electronic files may have previously been scanned from paper into electronic form (hence the term "blowback") and/or originated in native electronic form. In either event, somewhere along the way the files may have been converted into .pdf or .tif and/or endorsed with Bates numbers, privilege stamps, confidentiality redaction overlays, etc.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005).

Printing tiff or pdf images on paper. It is called blowback because originally paper documents may be scanned and then reprinted on paper from the scanned images.

BMP

A native file format of Windows for storing images called bitmaps.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Boolean Search

A search technique that utilizes Boolean Logic to connect individual keywords or phrases within a single query such as AND, OR, and NOT, within (w/5) , and NOT withinN (not w/5).

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

A Keyword Search in which the Keywords are combined using operators such as “AND,” “OR,” and “[BUT] NOT.” The result of a Boolean Search is precisely determined by the words contained in the Documents. (See also Bag of Words method.)

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The term "Boolean" refers to a system of logic developed by an early computer pioneer, George Boole. In Boolean searching, an "and" operator between two words results in a search for documents containing both of the words. An "or" operator between two words creates a search for documents containing either of the target words. A "not" operator between two words creates a search result containing the first word but excluding the second.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Applied Discovery's Glossary, http://www.lexisnexis.com/applieddiscovery/clientResources/glossary_B.asp

Source: RSI, Glossary.

A search type using Boolean logic operators between search terms that indicate a relationship between them. An "AND" operator between two words or other values (for example, "pear AND apple") means one is searching for documents containing both of the words or values, not just one of them. An "OR" operator between two words or other values (for example, "pear OR apple") means one is searching for documents containing either of the words.

Source: Ibis Consulting, Glossary.

Mathematical query language developed by English mathematician George Boole in the 19th century. Boolean searching of text is based on the underlying logic functions of various true/false statements. Common Boolean operators are “and,” “but not,” and “within.”

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A search for information using “AND,” “OR” and “NOT” commands, such as “Tom but not Jones” or “bankruptcy and trustee.”

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The use of the terms “AND,” “OR” and “NOT” in conducting searches. Used to widen or narrow the scope of a search.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Boot

The process whereby a computer automatically loads its startup software when it has been turned on. Also called “boot up,” the term derives from the phrase “to lift oneself up by one’s own bootstraps.”

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Box

A square graphic element on a form used to enter a single character, usually used in strings for entering constrained data.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

BPI

See: Bits Per Inch

BPS (Bits Per Second)

See: Bits Per Second (BPS)

Briefcase

A method to simplify the transport of a group of documents from one computer to another.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Bulk Coding

The process of Coding all members of a group of Documents (identified, for example, by Deduplication, Near-Deduplication, Email Threading, or Clustering) based on the review of only one or a few members of the group. Also referred to as Bulk Tagging.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Bulletin Board System

See: BBS (Bulletin Board System)

Bum

Slang for making (burning) a CD-ROM copy of data, whether it is music, software, or other data.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Burn CD burning

Burn

The process of creating a copy of information onto a CD-Rom or DVD.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

To record or write data on a CD or DVD.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Bum CD burning

Bus

A pathway between hardware devices, which may be internal, as is the case with components of a computer, or external, as is the case with computers in a network.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The "highway" which connects the various components of a computer system.

Source:Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Business Process Outsourcing

Business process outsourcing occurs when an organization turns over the management and optimization of a business function, such as accounts payable or purchasing, to a third party that conducts the activity based on a set of predetermined performance metrics.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Byte

A unit of measure consisting of eight bits that is the basic measurement of most computer data as multiples of the byte value. One million bytes are equivalent to a "megabyte" while one billion bytes is a "gigabyte."

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Eight bits.

Source:Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A computer word or a sequence of bits used as one unit, usually eight bits long. In word processing, a single character, such as a letter, is usually one byte in size.

Source:Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Eight bits. The ASCII standard to define letters, numbers and characters – maximum of 256. KB – Kilo-bytes, a thousand bytes (actually 2 10 or 1024 bytes). MB – Megabytes, a million bytes, (actually 2 20 or 1,024 KB or 1,048,576 bytes) GB – Gigabytes, a billion bytes (actually 2 30 or 1024 MB or 1,073,741,824 bytes).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Eight bits. A byte is a collection of bits used by computers to represent a character (i.e., "a", "1", or "&"). A "megabyte" is one million bytes or eight million bits or a "gigabyte" is one billion bytes or eight billion bits. 1 gigabyte = 1,000 megabytes. 1 terabyte = 1,000 gigabytes.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Standard unit of measure for computer storage. A byte is 8 bits (binary digits) and corresponds to about 1 English character.

See also:

Bit	MB - megabyte	TB - terabyte
KB - kilobyte	GB - gigabyte	PB - petabyte

EB - exabyte

Byte Level Deletion

Deletion is the process whereby data is removed from active files and other data storage structures on computers and rendered inaccessible except using special data recovery tools designed to recover deleted data. Deletion occurs in several levels on modern computer systems:

1. File level deletion: Deletion on the file level renders the file inaccessible to the operating system and normal application programs and marks the space occupied by the file's directory entry and contents as free space, available to reuse for data storage.
2. Record level deletion: Deletion on the record level occurs when a data structure, like a database table, contains multiple records; deletion at this level renders the record inaccessible to the database management system (DBMS) and usually marks the space occupied by the record as available for reuse by the DBMS, although in some cases the space is never reused until the database is compacted. Record level deletion is also characteristic of many e-mail systems.
3. Byte level deletion: Deletion at the byte level occurs when text or other information is deleted from the file content (such as the deletion of text from a word processing file); such deletion may render the deleted data inaccessible to the application intended to be used in processing the file, but may not actually remove the data from the file's content until a process such as compaction or rewriting of the file causes the deleted data to be overwritten.

Source: Merrill Corporation, Electronic Discovery Glossary.

Deletion is the process whereby data is removed from active files and other data storage structures on computers and rendered inaccessible except using special data recovery tools designed to recover deleted data.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Removing active files making them unavailable. Special data recovery tools can still retrieve these files.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

C

Cache

A form of high-speed memory used to temporarily store frequently accessed information; once the information is stored, it can be retrieved quickly from memory rather than from the hard drive.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A dedicated, high speed portion of computer memory which can be used for the temporary storage of frequently used data to make the application run faster (prevents having to constantly access the data from disk/tape storage).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A type a computer memory that temporarily stores frequently used information for quick access.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Caching

Of images: The temporary storage of image files on a hard disk for later migration to permanent storage, like an optical or CD jukebox.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

CAR

See: Computer Assisted Review (CAR)

Case De-Duplication

Retains only single copies of documents per case. For example, if an identical document resides with Mr. A, Mr. B and Mr. C, only the first occurrence of the file will be saved (Mr. A's). Contrast with custodian de-duplication and production de-duplication.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Basic de-duplication
Custodian de-duplication
De-duplication
Duplicate
Dynamic de-duplication
Global Deduplication
Horizontal Deduplication
Production de-duplication
Vertical Deduplication

Case Management System (CMS)

Software designed to regulate all law office functions performed with computers from one central application.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Case Search / Specifying Case

Specifying that the search must be case sensitive will match the exact case for all letters in the keyword and in the documents. For example, a case-sensitive search on Rose will match the name “Rose Jones” but it will not match the phrase “rose garden”.

Source: EDRM Search Glossary.

CCD (Charge Coupled Device)

A computer chip (with say 2048 cells) whose output is proportional to the light or color passed by it. Individual CCD's or arrays of these are used in scanners as a high-resolution, “digital camera" to "read" documents. These devices are micro-chip size and their resolutions run as high as 1000 pixels per inch.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

CCITT (Consultative Committee for International Telephone & Telegraphy)

Sets standards for phones, faxes, modems etc. The standard exists primarily for fax documents.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

CCITT Group 4

CCITT Group 4

A compression technique/format that reduces a file generally, about 5:1 over RLE and 40:1 over bitmap. For example, at a 300 bpi scan rate, the approximate storage requirements are: Size Raw RLE Group 4 A 1MB 200K 40K B 2MB 400K 75K C 4MB 820K 150K D 8MB 1.6MB 300K E 16MB 3.2MB 580K.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

CCITT

CD (Compact Disc or Compact Disk)

A 4 3/4" diameter device which can be read by a laser beam.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A removable optical disk that can be used to store documents or other data. CDs are available that can be both read and written using widely available CD "burners." It is common to transfer large amounts of data from one computer to another using CDs.

See also:

CD-R	DVD-ROM	Magnetic storage media
CD-ROM	Floppy disk	Media
CD-RW	Hard disk	Optical disk
Disc	Hard drive	Storage media
Disk	Jaz disk	WORM disk
Diskette	Laser disc	Zip disk
DVD	Magnetic disk	

CD Burning

The process of writing output to CD-ROMs or DVDs.

Source: Ibis Consulting, Glossary.

See also:

Bum	Burn

CD Publishing

An alternative to photocopying large volumes of paper documents. This method involves coupling image and text documents with viewer software on CDs. Sometimes search software is included on the CDs to enhance search capabilities.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

CD-R (C D-Recordable)

This is a CD that can be written (or recorded) only once. It can be copied to distribute a large amount of data. CD-Rs can be read on any CD-ROM drive whether on a standalone computer or network system. This makes interchange between systems easier.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Often also used as an acronym for CD-ROM's that can be written more than once. The succeeding writings must utilize unused sections of the original, with a library o directory of the total use. Optical storage technology using formats compatible with CD-ROM's. CD-ROM discs

must be "pre-mastered" to insure that the data is correctly formatted. Using a "double speed" recorder, it takes about a half hour to burn a complete 650MB disc.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

CD	DVD-ROM	Magnetic storage media
CD-ROM	Floppy disk	Media
CD-RW	Hard disk	Optical disk
Disc	Hard drive	Storage media
Disk	Jaz disk	WORM disk
Diskette	Laser disc	Zip disk
DVD	Magnetic disk	

CD-Recordable

See: CD-R (C D-Recordable)

CD-ROM (Computer Disk Read Only Memory)

Optical disk storage using the same technology as audio CDs. A computer can read a CD-ROM disc but cannot write on it. Typically used to distribute large amounts of textual information, since one CD-ROM holds about 650 MB of data, or approximately 15,000 pages of text.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A type of high density optical disk with a 4" diameter and a 650MB capacity. The information (1's or 0's) is permanently etched by a laser into the surface of the disk and read by a laser beam. The ISO 9660 standard defines how a CD-ROM is written for computer interface. It is not rewritable. It is legally accepted and written on a single-side.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Written on a large scale and not on a standard computer CD burner (CD writer), they are an optical disk storage media popular for storing computer files as well as digitally recorded music.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Data storage medium that uses compact discs to store about 1,500 floppy discs' worth of data.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

CD
CD-R
CD-RW
Disc
Disk
Diskette
DVD
DVD-ROM
Floppy disk
Hard disk
Hard drive
Jaz disk
Laser disc
Magnetic disk
Magnetic storage media
Media
Optical disk
Storage media
WORM disk
Zip disk

CD-ROM Drive

A computer drive that reads compact discs.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

CD-RW (Compact Disc Re-Writable)

See also:

CD
CD-R
CD-ROM
Disc
Disk
Diskette
DVD
DVD-ROM
Floppy disk
Hard disk
Hard drive
Jaz disk
Laser disc
Magnetic disk
Magnetic storage media
Media
Optical disk
Storage media
WORM disk
Zip disk

CDMA (Code-Division Multiple Access)

An emerging wireless communication technology for all digital voice and data networks.

Source: RenewData, Glossary (10/5/2005).

CDPD (Cellular Digital Packet Data)

A data communication standard which uses the unused capacity (bandwidth) of cellular voice providers.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

CDR (Computer Disk Recorder)

The machine that actually “burns” information onto a CD.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Cellular Digital Packet Data

See: CDPD (Cellular Digital Packet Data)

Central Processing Unit (CPU)

The “brain” of the computer. In a PC, the CPU is contained on a single microprocessor chip and performs all logical operations to run programs and solve problems.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The portion of a computer which performs most of the logical and arithmetic functions.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Centronics Interface

A parallel interface standard for connecting printers and other devices to computers. Pioneered by the Centronics Inc., a printer manufacturer in New Hampshire. Uses a 36 pin connector. See SPP.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Certified Forensic Examiner

A person holding one of a number of commonly recognized certifications in the field. Due to a lack of industry wide certifications it is critical to research the certifications and any requirements within your state or jurisdiction.

Source: EDRM Collection Standards

CGA (Color Graphics Adapter)

Short for Color Graphics Adapter, CGA was an early IBM video adapter that replaced monochrome and was first introduced in 1981. CGA has the highest resolution of 640 x 200, color depth of 4-bit, and supports 16 colors (24 = 16).

Source: Computer Hope, CGA definition, http://www.computerhope.com/jargon/c/cga.htm

The Color Graphics Adapter (CGA), originally also called the Color/Graphics Adapter or IBM Color/Graphics Monitor Adapter, introduced in 1981, was IBM's first graphics card and first

color display card for the IBM PC. For this reason, it also became that computer's first color computer display standard.

Source: Wikipedia, Color Graphics Adapter, https://en.wikipedia.org/wiki/Color_Graphics_Adapter

See also:

VGA

Chain of Custody

All information on a file's travels from its original creation version to its final production version. A detailed account of the location of each document/file from the beginning of a project until the end. A sound chain of custody verifies that you have not altered information either in the copying process or during analysis. If you cannot show the chain of custody, you may have a difficult time disproving that outside influences might have tampered with the data. A chain of custody failure — i.e., the mishandling of electronic evidence (even fully recovered files) — can cause a litigation defeat.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Feldman, The Essentials of Computer Discovery, Computer Forensics Inc. (1/1/2001), http://www.forensics.com/pdf/Essentials_of_Discovery.pdf#page=12

A process used to maintain and document the chronological history of the handling of electronic evidence. A chain of custody ensures that the data presented is "as originally acquired" and has not been altered prior to admission into evidence. Some providers maintain an electronic chain-of-custody link between all electronic data and its original physical media throughout the production process.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

An accounting of the control (custody) of real evidence at all times until the moment it is offered in evidence. Chain of custody helps to show that the evidence being offered has not been tampered with and is authentic. Chain of custody is important for electronic evidence because it can be easily altered.

Source: Ibis Consulting, Glossary.

Chain of custody refers to the chronological documentation and/or paper trail showing the seizure, custody, control, transfer, analysis, and disposition of evidence, physical or electronic. Because evidence can be used in court to convict persons of crimes, it must be handled in a scrupulously careful manner to avoid later allegations of tampering or misconduct, which can compromise the case of the prosecution toward acquittal or become grounds for overturning a guilty verdict upon appeal. The idea behind recording the chain of custody is to establish that the alleged evidence is in fact related to the alleged crime - rather than, for example, having been planted fraudulently to make someone appear guilty.

Establishing chain of custody is especially important when the evidence consists of fungible goods. In practice, this most often applies to illegal drugs that have been seized by law enforcement personnel. In such cases, the defendant may disclaim any knowledge of possession of the controlled substance in question. Accordingly, the chain of custody documentation and testimony is presented by the prosecution to establish that the substance in evidence was in fact in the possession of the defendant.

An identifiable person must always have the physical custody of a piece of evidence. In practice, this means that a police officer or detective will take charge of a piece of evidence, document its collection, and hand it over to an evidence clerk for storage in a secure place. These transactions, and every succeeding transaction between the collection of the evidence and its appearance in court, should be completely documented chronologically in order to withstand legal challenges to the authenticity of the evidence. Documentation should include the conditions under which the evidence is gathered, the identity of all evidence handlers, duration of evidence custody, security conditions while handling or storing the evidence, and the manner in which evidence is transferred to subsequent custodians each time such a transfer occurs (along with the signatures of persons involved at each step).

Source: EDRM Presentation Guide.

See also:

Chain of evidence

Forensically sound procedures

Chain of Evidence

The "sequencing" of the chain of evidence follows this order: identification and collection; analysis; storage; preservation; transportation; presentation in court; return to owner. The chain of evidence shows: who obtained the evidence; where and when the evidence was obtained; who secured the evidence; who had control or possession of the evidence. 1

See also:

Chain of custody

Forensically sound procedures

Character Encoding

Electronic data is represented as sequences of bits, or numbers. Each alphabet or script used in a language is mapped to a unique numeric value. This is referred to as character encoding. See also Unicode.

Source: EDRM Search Glossary.

Character Treatment

The use of all caps or another standard form of treating letters in a coding project.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Characters Per Inch (CPI)

Charge Coupled Device

See: CCD (Charge Coupled Device)

Chip

A piece of silicon containing electronic circuits that perform computing functions processed by the chip. The chip is mounted onto a socket that has a number of projecting pins and fits into a receptacle on the motherboard.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

CIE (Commission International de l'Eclairage)

The international commission on color matching and illumination systems.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Cine-Mode

Data recorded on a film strip such that it can be read by a human when held vertically.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Cinepak

A compression algorithm, see MPEG.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

CITIS (Contractor Integrated Technical Information Service)

The Department of Defense now requires contractors to have an electronic document image and management system.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Civil Procedure Rules (CPR)

The process formerly known as discovery by which documents are exchanged between parties in litigation in England and Wales. Technically the process has three phases - a) disclosure - making it known that the documents exist by providing the other party with a list, b) inspection - allowing the other party to look at the documents c) the provision of copies. Typically all three

phases are dealt with together. Unlike the US, production of documents is initially driven by "push" i.e. there is an obligation on a party to disclose their documents which are material to the case. For full details see CPR 31 and PD 31.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

Classical, Gaussian, or Normal Calculator / Classical, Gaussian, or Normal Estimation

A method of calculating Confidence Intervals based on the assumption that the quantities to be measured follow a Gaussian (Normal) Distribution. This method is most commonly taught in introductory statistics courses, but yields inaccurate Confidence Intervals when the Prevalence of items with the characteristic being measured is law. (C.f. Binomial Calculator / Binomial Estimation.)

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Classifier / Classification / Classified / Classify

To arrange or designate according to categorization such as potentially responsive or privileged versus non-responsive or not-privileged.

Source: EDRM Search Glossary.

An Algorithm that Labels items as to whether or not they have a particular property; the act of Labeling items as to whether or not they have a particular property. In Technology-Assisted Review, Classifiers are commonly used to Label Documents as Responsive or Non-Responsive.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Classify / Classification

To arrange or designate according to categorization such as potentially responsive or privileged versus non-responsive or not-privileged.

Source: EDRM Search Guide Glossary.

Clean Install

A clean install is a software installation in which any previous version is eradicated. The alternative to a clean install is an upgrade, in which elements of a previous version remain.

Source: http://searchitchannel.techtarget.com/definition/clean-install.

Client/Server Network

A computer system functionally distributed across several nodes on a network, sometimes called a distributed application. The basic theory is that the various components of the system

can be tailored to perform specific functions, hopefully for the good of the entire network. Client/Server systems are also typified by a high degree of parallel processing across distributed nodes. Usually the clients are individual PC's connected to server(s) which act as central storehouses and "traffic cops" for information and applications.

Compare with file-sharing applications, where all searches occur on the workstation, while the document database resides on the server. With client-server architecture, CPU intensive processes (such as searching and indexing) are completed on the server, while image viewing and OCR occur on the client. File-sharing applications are easier to develop, but they tend to generate tremendous network data traffic in document imaging applications. They also expose the database to corruption through workstation interruptions. Client-server applications are harder to develop, but dramatically reduce network data traffic and insulate the database from workstation interruptions.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

LAN - local area network	Peer-to-peer network	WAN - wide area network
MAN - metropolitan area network	SAN - storage area network	
Network	Stand alone computer	

Clock Speed

The speed with which the computer processes information. PC clock speed is measured in megahertz, e.g., 60 megahertz.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Cloud Computing

Massive and shared computing resources where each user obtains the computing and storage resources they need from a common pool, usually owned by a third-party data service and housed in their data center. Cloud computing is, in some ways, reminiscent of time sharing main-frame computing, where a few big companies controlled to computational resources for most users.

Source: Herb Roitblat, Search 2020: The Glossary.

Cluster

In operating systems that use a file allocation table (FAT) architecture, the smallest unit of storage space required for data written to a drive. Also called an allocation unit.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

The smallest unit of storage space required for computer data to be written to a drive. Sometimes called an allocation unit.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Cluster (File): The smallest unit of storage space that can be allocated to store a file on operating systems that use a file allocation table (FAT) architecture. Windows and DOS organize hard discs based on clusters (also known as allocation units), which consist of one or more contiguous sectors. Discs using smaller cluster sizes waste less space and store information more efficiently.

Cluster (System): A collection of individual computers that appear as a single logical unit. Also referred to as matrix or grid systems.

Clustering

An Unsupervised Learning method in which Documents are segregated into categories or groups so that the Documents in any group are more similar to one another than to those in other groups. Clustering involves no human intervention, and the resulting categories may or may not reflect distinctions that are valuable for the purpose of a search or review effort. 1

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Grouping documents or other objects by similarity. The similarity between two documents in a cluster is greater than the similarity of documents in two different clusters.

Source: Herb Roitblat, Search 2020: The Glossary.

Source: Herb Roitblat, Predictive Coding Glossary

CMS

See: Case Management System (CMS)

CMYK (Cyan, Magenta, Yellow and Black)

A subtractive method used in four color printing and desktop publishing.

Sourec: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Co-Processor

An additional processor, which performs specific tasks while the main processor runs the primary functions of the system. A math co-processor, for example, performs arithmetic operations to take that burden off the main processor resulting in faster operations.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Code / Coded / Coding

The action of Labeling a Document as Relevant or Non-Relevant, or the set of Labels resulting from that action. Sometimes interpreted narrowly to include only the result(s) of a Manual Review effort; sometimes interpreted more broadly to include automated or semi-automated Labeling efforts. Coding is generally the term used in the legal industry; Labeling is the equivalent term in Information Retrieval.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Code-Division Multiple Access

See: CDMA (Code-Division Multiple Access)

Coder

An individual assigned to input information from documents into a document database.

Sourec: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Coding

An automated or human process where text content is compared to pre-determined codes, and the results of those comparisons are logged. Coding usually identifies names, dates, and relevant terms or phrases. Coding may be structured, i.e., a Yes/No option as to an issue or the selection of one of the finite number of choices, or unstructured, i.e., a narrative comment about a document. Coding may be objective, i.e., the name of the sender or the date, or subjective, i.e., evaluation as to the documents.

Source: Ibis Consulting, Glossary.

A means of capturing specific, standardized data from a collection of documents and creating a database linking the data to the images. The term “coding” is generally used in the legal and medical markets. It is similar to “indexing” in the commercial marketplace.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Document coding is the process of capturing case-relevant information (i.e. author, date authored, date sent, recipient, date opened, etc.) from a paper document.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Bibliographic Coding	Level coding	Taxonomic coding
Indexing	Objective coding	Verbatim coding
Issue Code	Subjective coding	
Issue coding	Tag	

Coding Manual

A set of instructions provided to coders that includes a description of the project, subject codes, and rules for data conformance and consistency.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Cognitive Computing

A style of computing intended to mimic the way the human mind works. It is intended to address the kinds of problems where human judgment has previously been required, problems involving high amounts of ambiguity and uncertainty. Cognitive computing typically involves sophisticated forms of natural language processing and automatic reasoning.

Source: Herb Roitblat, Search 2020: The Glossary.

COLD (Computer Output to Laser Disk)

The computer system contains files of ASCII data (from input or application programs) or bit-mapped files previously scanned from microfilm documents or pictures. These output files are compressed by a factor of 5-20:1 from the original documents and stored on WORM optical/laser disks. The stored data is then available to all on the network. Generally, the format of these databases are compatible with SQL and imaging formats.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Collection

A group of documents. These can be documents gathered for a particular matter or purpose. Information retrieval scientists tend use several well-known document collections (e.g., RCV1) for testing and comparison purposes.

Source: Herb Roitblat, Predictive Coding Glossary.

Collection Phase

Gathering ESI for further use in the e-discovery process (from EDRM Stages web page).

Source: EDRM Stages

Corresponds to UTBMS Code L620-L629. Collection/Recovery, Media Costs, Media/ESI Transfer, Receipt, Inventory, Quality Assurance and Control.

Source: EDRM Metrics Glossary

Color Graphics Adapter

See: CGA (Color Graphics Adapter)

COM (Computer Output to Microfilm)

The computer converts and stores data directly on microfilm/fiche from a variety of available inputs. This older technology is cheaper and more convenient than paper, but one of the most difficult to use in actually storing and retrieving the data.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Comb

A series of boxes with their tops missing. Tick marks guide text entry. Used in forms processing rather than boxes.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Combined Word Search

A word search that combines synonym, proximity, and/or Boolean searches.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming

Synonym search
Topical search
Term search
Weighted relevance search
Wildcard search

Comic Mode

Human-readable data, recorded on a strip of film which can be read when the film is moved horizontally to the reader.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Commission International de l'Eclairage

See: CIE (Commission International de l'Eclairage)

Common User Interface (CUI)

IBM' s answer to the Apple Macintosh, it is a standard for menus and windows developed by IBM.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Compact Disc

See: CD (Compact Disc or Compact Disk)

Compact Disc Re-Writable

See: CD-RW (Compact Disc Re-Writable)

Compact Disc Recordable

See: CD-R (C D-Recordable)

Compact Disk

See: CD (Compact Disc or Compact Disk)

Compatibility

A characteristic of a computer or software by which data prepared in another computer or software can be processed.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The interchangeability of computer components, either hardware or software.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Compliance Search

Searching for the purposes of identification of specified relevant information in response to a discovery request. A compliance search should be paired with a methodology search as Ad-Hoc or Iterative searching.

Source: EDRM Search Glossary.

Comply

Used in the context of a discovery request. When one complies with a discovery request, it is through a production (of either a witness or document).

Source: Ibis Consulting, Glossary.

Composite Video

A video stream that combines red, green, blue and synchronization signals into one so it only requires one connector. Composite video is used by most televisions and VCR's. Separate luminosity and color signals that provide the highest possible signal quality. Distinct from video standards such as NTSC or PAL.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Compound Document

A document that contains linked or embedded objects as well as its own data.

Source: Ibis Consulting, Glossary.

Compression

A technology for storing data in fewer bits, it makes data smaller so less disk space is needed to represent the same information. Compression programs like WinZip and UNIX compress are valuable to network users because they save both time and bandwidth. Data compression is also widely used in backup utilities, spreadsheet applications, and database management systems.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

A technology for storing data in fewer bits, it makes data smaller so less disk space is needed to represent the same information. Data compression is widely used to backup utilities, spreadsheet applications, and database management systems. Compressed files must be decompressed in order to be useable.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A technology that reduces the size of a file.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Any method which reduces the amount of data necessary to transmit information from one point to another. Compression generally eliminates redundant information and/or predicts where changes will occur. "Lossless" compression techniques totally preserve the integrity of the input. "Lossy" methods disregard some of the originals. The ratio of the file sizes of a compressed file to an uncompressed file, e.g., with a 20:1 compression ratio, an uncompressed file of 1MB is compressed to 50KB.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A technology that reduces the size of a file. Compression programs are valuable to network users because they help save both time and bandwidth.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Computer

Includes but is not limited to network servers, desktops, laptops, notebook computers, employees' home computers, mainframes, PDAs (personal digital assistants, such as PalmPilot, Cassiopeia, HP Jornada and other such handheld computing devices), digital cell phones and pagers.

Source: RSI, Glossary.

See also:

File server
Laptop computer
Microcomputer
Minicomputer
Notebook computer
Personal computer
Workstation

Computer Assisted Review (CAR)

Any of a number of technologies that use computers to facilitate the review of documents for discovery.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

CAR
Predictive Coding
TAR

Computer Disk Read Only Memory

See: CD-ROM (Computer Disk Read Only Memory)

Computer Disk Recorder

See: CDR (Computer Disk Recorder)

Computer Evidence

Computer evidence is rather unique when compared to other forms of more traditional documentary evidence. Unlike paper documentation, computer evidence is extremely fragile and it occurs in the form of an identical copy of a specific document that is stored in a computer file. In addition, the legal "best evidence" rules differ for the processing of computer evidence. However, there is the potential for unauthorized copies to be made of important computer files without leaving behind a trace that a copy was made. Computer evidence is not limited to data stored in computer files, rather most relevant computer evidence is uncovered in uncommonly known locations. For example, on Microsoft Windows and Windows NT-based computer systems, large quantifies of evidence can be found in the Windows swap files or Page Files. In addition computer evidence can also be uncovered in file slack and unallocated file space.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

See also:

Computer forensics
Computer investigations
Discovery
Electronic discovery / e-discovery
Electronic evidence
Forensic analysis
Forensics
Mirroring

Computer Forensics

Computer forensics is the use of specialized techniques for recovery, authentication and analysis of electronic data when a case involves issues relating to reconstruction of computer usage, examination of residual data, and authentication of data by technical analysis or explanation of technical features of data and computer usage. Computer forensics requires specialized expertise that goes beyond normal data collection and preservation techniques available to end-users or system support personnel.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Source: Merrill Corporation, Electronic Discovery Glossary.

Similar to all forms of forensic science, computer forensics is comprised of the application of the law to computer science. Computer forensics deals with the preservation, identification, extraction, and documentation of computer evidence. Like many other forensic sciences, computer forensics involves the use of sophisticated technological tools and procedures that must be followed to guarantee the accuracy of the preservation of evidence and the accuracy of results concerning computer evidence processing.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The use of specialized techniques for recovery, authentication, and analysis of computer data, typically of data which may have been deleted or destroyed.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Computer evidence
Computer investigations
Discovery
Electronic discovery / e-discovery
Electronic evidence
Forensic analysis
Forensics
Mirroring

Computer Investigations

Computer crimes are specifically defined by federal and/or state statutes and any computer documentary evidence utilized during a computer investigation may include computer data stored on floppy diskettes, zip disks, CDs and computer hard disk drives. The evidence necessary to prove computer-related crimes can potentially be located on one or more computer hard disk drives in various geographic locations. This evidence can reside on computer storage media as bytes of data in the form of computer files and ambient data, however, ambient data is usually unknown to most computer users and is therefore often very useful to computer forensic investigators. Computer investigations rely upon evidence stored as data and the timeline of dates and times that files were created, modified, and/or last accessed by a computer user. Timelines of activities can be essential when multiple computers and individuals are involved in the commission of a crime. In addition, computer investigations generally involve the review of Internet log files to determine Internet account abuses. Using computer forensic procedures, processes, and tools, computer forensics investigators can identify passwords, network logons, Internet activity, and fragments of email messages that were dumped from computer memory during past Windows work sessions.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

See also:

Computer evidence
Computer forensics
Discovery
Electronic discovery / e-discovery
Electronic evidence
Forensic analysis
Forensics
Mirroring

Computer Output to Laser Disk

See: COLD (Computer Output to Laser Disk)

Computer Output to Microfilm

See: COM (Computer Output to Microfilm)

Concept Search

A search technique that provides words which are similar in concept to a query word. A concept search will return documents that relate to the same concept as the query word, regardless of whether the query word exists in the search results documents. Concept searches can be

implemented as a simple thesaurus match, or by using sophisticated statistical analysis methods. Effectiveness of concept search in an e-discovery project depends greatly on the type of algorithm used and its implementation.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

An industry-specific term generally used to describe Keyword Expansion techniques, which allow search methods to return Documents beyond those that would be returned by a simple Keyword or Boolean Search. Methods range from simple techniques such as Stemming, Thesaurus Expansion, and Ontology search, through statistical Algorithms such as Latent Semantic Indexing.

Source: The Grossman-Cormack Glossary of Technology Assisted Review (Version 1.02, Nov. 2102).

Maps relationships between each word and every other word in large sets of documents and then associates words based on the context in which they are used. Two techniques can be used to perform concept searches: the use of a manually constructed thesaurus which relates certain words to others or semantic indexing, a fully automated method to show associations among words based, in part, on statistical analysis of the occurrence of proximity of certain words to others.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Also called "thesaurus" or "related" searching; sometimes called "synonym searching." Searches that provide other words similar or close in meaning to the primary word.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Confidence Interval

As part of a Statistical Estimate, a range of values estimated to contain the true value, with a particular Confidence Level.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The expected range of results. If you drew repeated samples from the same population, you would expect the result to be within the confidence interval about the proportion of times given by the confidence level. For example, in an election poll, the difference in the proportion of people favoring each candidate is described as being within a range of, say, plus or minus 5%. All other things being equal, the smaller the confidence interval, the larger the sample size needs to be. Said another way, the larger the sample size, the smaller the confidence interval.

Source: Herb Roitblat, Search 2020: The Glossary.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

Margin of Error

Confidence Level

As part of a Statistical Estimate, the chance that a Confidence Interval derived from a Random Sample will include the true value. For example, “95% Confidence” means that if one were to draw 100 independent Random Samples of the same size, and compute the Confidence Interval from each Sample, about 95 of the 100 Confidence Intervals would contain the true value. It is important to note that the Confidence Level is not the Probability that the true value is contained in any particular Confidence Interval; it is the Probability that the method of estimation will yield a Confidence Interval that contains the true value.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

How often we would achieve a similar result if we repeated the same process many times. If we did the same kind of test from the same population more than once, the confidence level would tell us how often we would get a result that is within a certain range (the confidence interval) of the true value for the population. Most scientific studies employ a minimum confidence level of 0.95, meaning that 95 percent of the time when you repeated the experiment you would find a similar result. The higher the confidence level the larger the sample size that is required. Technically, it is the proportion of times when the true population value would be included within the confidence interval.

Source: Herb Roitblat, Search 2020: The Glossary.

Source: Herb Roitblat, Predictive Coding Glossary.

Config.sys

A DOS configuration file, which is used when the computer boots, to load specific device drivers to run hardware or software components.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Confusion Matrix

A two-by-two table listing values for the number of True Negatives (TN), False Negatives (FN), True Positives (TP), and False Positives (FP) resulting from a search or review effort. As shown below, all of the standard evaluation measures are algebraic combinations of the four values in the Confusion Matrix. Also referred to as a Contingency Table. An example of a Confusion Matrix (or Contingency Table) is provided immediately below.

	Coded Relevant	Coded Non-Relevant
Truly Relevant	True Positives (TP)	False Negatives (FN)
Truly Non-Relevant	False Positives (FP)	True Negatives (TN)

Accuracy = 100% – Error = (TP + TN) / (TP + TN + FP + FN)

Elusion = 100% – Negative Predictive Value = FN / (FN + TN)

Error = 100% – Accuracy = (FP + FN) / (TP + TN + FP + FN)

Fallout = False Positive Rate = 100% – True Negative Rate = FP / (FP+TN)

False Negative Rate = 100% – True Positive Rate = FN / (FN+TP)

Negative Predictive Value = 100% – Elusion = TN / (TN + FN)

Precision = Positive Predictive Value = TP / (TP + FP)

Prevalence = Yield = Richness = (TP + FN) / (TP + TN + FP + FN)

Recall = True Positive Rate = Sensitivity = TP / (TP+FN)

True Negative Rate = Specificity = TN / (TN + FP)

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Consultative Committee for International Telephone & Telegraphy

See: CCITT (Consultative Committee for International Telephone & Telegraphy)

Container

An application or object that contains other files or objects which can be represented as files. A container might be an archive or a compound document with an embedded or linked object.

Source: Ibis Consulting, Glossary.

See also:

EML	NSF	Single-mail archive
Mail container	OST	Single-mail container
Mailbox	PST	SMTP
MSG	RFC compliant email	
Multi-mail container	RFC822	

Contingency Table

A table of the four response states in a categorization task. The rows of the table may correspond to the correct or true category values and the columns may correspond to the choices made by system. For example, the top row may be the truly positive category (e.g. truly responsive documents) and the second row may be the truly negative category (e.g., truly non-responsive documents). The columns then represent the positive decisions made by the system (e.g., putatively responsive) and the negative decisions made by the system (e.g., putatively non-responsive). The entries in these cells are the counts of documents corresponding to each response state (e.g., true positives, false negatives, false positives, true negatives). Contingency tables are often displayed along with the totals for each row and for each column. Sometimes the rows and columns are reversed, so the columns reflect the true values and the rows reflect the choices.

Source: Herb Roitblat, Predictive Coding Glossary.

Continuous Tone

An image (e.g. a photograph) which has all the values of gray from white to black.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Contractor Integrated Technical Information Service

See: CITIS (Contractor Integrated Technical Information Service)

Control Number

See: Bates Number

Control Number Prefix

A project-specific, client specification in the form of an alphanumeric prefix that precedes a project's control number (the digital equivalent of a Bates number). Also called "control number prefix."

Source: Ibis Consulting, Glossary.

See also:

Bates number | Bates stamp | Document number

Bates prefix | Bates stamping

Control Set

A Random Sample of Documents Coded at the outset of a search or review process, that is separate from and independent of the Training Set. Control Sets are used in some Technology-Assisted Review processes. They are typically used to measure the effectiveness of the Machine Learning Algorithm at various stages of training, and to determine when training may cease.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Convergence

Where the RGB signals "converge" on a single pixel. That pixel should be white at full brightness of the RGB components.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Cookie

A data set that a web site server gives to a browser the first time a user visits a site, updated with each return visit. The remote server saves cookie data about a user as text files stored in Netscape or MS Internet Explorer system folders. Cookies may contain user or session specific data such as user name, date of visit, statistic and anything that server knows about remote user. Cookies may be updated one or more times each visit, or only once.

Source: Ibis Consulting, Glossary

Holds information on the times and dates a user has visited websites. Other information can also be saved to your hard drive in these text files, including information about online purchases, validation information about the user for "Members Only" websites, etc.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Small data file written to a user’s hard drive by a web server which contains information the web site uses to identify the user in subsequent visits.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Small data files written to a user's hard drive by a Web server. These files contain specific information that identifies users (i.e., passwords and lists of pages visited).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Copy/Paste

To copy a piece of data to a temporary location and then make a new copy of the object in a new location. This is usually done by clicking the right mouse button while holding the mouse cursor over the relevant file and then clicking "copy" from the menu that appears. The mouse pointer is then moved to the destination location, a right mouse click brings up the same function menu and "paste" is selected to copy the file(s) to the new location.

Source: EDRM Collection Standards

Copyee Field

A data field used to record the names of individuals and/or business entities who received a copy of a document, when the name is not otherwise recorded in the addressee or recipient field.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Corpus

A collection of objects, typically documents that are the subject of analysis, machine learning, or categorization.

Source: Herb Roitblat, Search 2020: The Glossary.

Corrupt File

A file with deteriorated data as a result of some external agent. Hazards to data integrity include not only computer-based problems such as viruses, hardware or software incompatibilities, flaws, or failures. Also environmental threats such as power outages, dust, water, and extreme temperatures can cause hardware failure that adversely affect file data.

Source: Ibis Consulting, Glossary.

Cost

Cost refers to the measurable dollars associated with each identifiable task, activity or action. Cost is also a variable element of Time and Volume. Cost may also include discrete elements that may be independent of Time or Volume.

Source: EDRM Metrics Glossary

Coverage Bias

Coverage Bias can occur if the samples are not representative of the population due to the methodology used. In e-discovery, such coverage bias occurs when large portions of ESI get excluded from based on meta-data or type of ESI. As an example, Patent Litigation may require sampling technical documents in their source form, and care should be taken to include these documents in the sample selection process.

Source: EDRM Search Glossary.

CPI

See: Characters Per Inch (CPI)

CPR

See: Civil Procedure Rules (CPR)

CPU

See: Central Processing Unit (CPU)

CRC (Cyclical Redundancy Checking)

Used in data communications to create a checksum character (hexadecimal) at the end of a data block.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Crime Scene Reconstruction

Crime scene reconstruction is the use of scientific methods, physical evidence, deductive reasoning, and their interrelationships to gain explicit knowledge of the series of events that may have led up to the crime and what exactly happened at a specific crime scene. It is a disciplined and principled approach towards objectively understanding a crime scene. Crime reconstruction helps interpret physical evidence. It is an aid to help formulate a hypothesis and arrive at a conclusion about a certain crime. Forensic specialists all come together with their different forms of evidence such as photos, sketches, and other useful things gathered from the crime scene to paint a vivid picture which makes it possible to retrace a crime that took place. Using evidence found at a proper crime scene you can reconstruct what happened and possibly find more clues.

When focusing on other types of forensics, there are three areas of importance in finding the answers and determining the components of a crime scene: (1) specific incident reconstruction, which deals with traffic accidents, bombings, homicides, and things of that nature; (2) event reconstruction, which analyzes connections, sequence, and identity; and the most important component, (3) physical evidence reconstruction, which focuses on firearms, blood, glass, and other objects that can be stripped for DNA.

Source: EDRM Presentation Guide.

Cross-Reference Field

A data field used to record information that is cross-referenced to the specific document record. May be used to cross-reference:

1. Parent documents with attachments;
2. Separate text pages to one another; and
3. Documents with different identifying numbers.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Crossover Trial

An Experimental Design for comparing two search or review processes using the same Document Collection and Information Need, in which one process is applied first, followed by the second, and then the results of the two efforts are compared. (Cf. Parallel Trial.)

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

CUI

See: Common User Interface (CUI)

Culling

The practice of narrowing a larger data set to a smaller data set for the purposes of review, based on objective criteria (such as file types or date restrictors), or subjective criteria (such as Keyword Search Terms). Documents that do not match the criteria are excluded from the search and from further review.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Cursor

A symbol used on the computer screen in DOS systems to show where data are to be entered.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Custodian

A common element within each e-discovery Phase which refers to the individual(s) responsible for data types or repositories for a given entity. Individuals in possession of data that is potentially relevant to a case.

Source: EDRM Metrics Glossary

See also:

Data custodian

Custodian De-Duplication

Culls a document if multiple copies of that document reside within the same custodian's data set. For example, if Mr. A and Mr. B each have a copy of a specific document, and Mr. C has two copies, the system will maintain one copy each for Mr. A, Mr. B, and Mr. C. Contrast with case de-duplication and production de-duplication.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Basic de-duplication

Case de-duplication

De-duplication

Duplicate

Dynamic de-duplication

Global Deduplication

Horizontal Deduplication

Production de-duplication Vertical Deduplication

Custodian Search

Custodian search is a common form of constraining search results. To search based on a custodian, the metadata search using the metadata name "Custodian" can be used. Custodian search may rely on assigning custodians to collected data during the Identification Phase so that searching doesn't miss out on custodians. For example, instant messages with buddy-names may be missed if the search term is specified as last-name/first-name or as email addresses.

Source: EDRM Search Glossary.

Custodians

A common element within each e-discovery Phase which refers to the individual(s) responsible for data types or repositories for a given entity. Individuals in possession of data that is potentially relevant to a case.

Source: EDRM Metrics Glossary

See also:

Data custodian

Customer-Added Metadata

Data or work product created by a user while reviewing a document. For example, annotation text of a document or subjective coding information. Contrast with vendor-added metadata.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Document metadata
Email metadata
Extrinsic data
File parameters
File system metadata
File-specific metadata
General metadata
Metadata
Vendor-added metadata

Customized Data Field

A specially named and defined data field in a database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Customized Field Definition

The process of defining the characteristics of customized data fields in a database, including field structure (date, text, or integer field), field size (number of characters), multiple values (more than one name or code in a field), and field name.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Cut and Paste

To highlight a block of text then move or copy it, either to another area of the same document or to a completely separate document.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Cutoff

A given score or rank in a Prioritized list, resulting from a Relevance Ranking search or Machine Learning Algorithm, such that the Documents above the Cutoff are deemed to be Relevant and Documents below the Cutoff are deemed to be Non-Relevant. In general, a higher Cutoff will yield higher Precision and lower Recall, while a lower Cutoff will yield lower Precision and higher Recall. Also referred to as a Threshold.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Cyan

A colored ink. Reflects blue & green & absorbs red.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Cyan, Magenta, Yellow and Black

See: CMYK (Cyan, Magenta, Yellow and Black)

Cyclical Redundancy Checking

See: CRC (Cyclical Redundancy Checking)

D

Da Silva Moore

Da Silva Moore v. Publicis Groupe, Case No. 11 Civ. 1279 (ALC) (AJP), 2012 WL 607412 (S.D.N.Y. Feb. 24, 2012), aff'd 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012). The first federal case to recognize Computer Assisted Review as "an acceptable way to search for relevant ESI in appropriate cases." The opinion was written by Magistrate Judge Andrew J. Peck and affirmed by District Judge Andrew L. Carter.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

DAC (Digital to Analog Converter)

Changes digital numbers to an electrical waveform.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

DAT (Digital Audio Tape)

Although generally used for audio, a DAT (120 meters long) can hold up to 10 gigabytes if used for digital data storage. Has the disadvantage of being a serial, rather than a random access device.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Used as a storage medium in some backup systems.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Records audio signals onto tape in a digital format. May also be used as a backup tape in some systems.

See also:

Backup
Backup tape
Data extraction
Digital audio tape
Disaster recovery tape
DLT - digital linear tape
Magnetic storage media
Media
QIC - quarter inch cartridge
Tape

Data

Numbers, characters, images, or other method of recording, in a form which can be assessed by a human or (especially) input into a computer, stored and processed there, or transmitted on some digital channel.

Source: Ibis Consulting, Glossary.

Any information stored on a computer.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

A general phrase for all information (facts, numbers, letters, graphics, etc.) that can be processed by a computer.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Information stored on the computer system and used by applications to accomplish tasks.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Data Attribute

A data attribute is a characteristic of data that sets it apart from other data, such as location, length, or type. The term attribute is sometimes used synonymously with “data element” or “property.”

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Data Center

A secure site used to house computer applications for use by one or more clients. Usually includes a higher level of security, power supply (generators, back-up switches, etc.) and telecommunications than is found in a standard computer room.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Data Custodian

Person having administrative control of a document or electronic file; for example, the data custodian of an email is the owner of the mailbox which contains the message.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Custodian

Dark Data

Data that are stored, perhaps in a data lake, in the hope that someday it might be useful to the organization. Dark data is typically big data that are unstructured, unanalyzed, uncategorized, and, most importantly, unused for any valuable business activity. Hoarded data, also called dusty data.

Source: Herb Roitblat, Search 2020: The Glossary.

Data Entry

The process of entering information into a database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Data Extraction

The process of removing files and meta-data from backup tapes.

Source: RenewData, Glossary (10/5/2005).

The process of restoring files and meta-data from backup tapes in order to make them accessible.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The process of pulling information out of either hard copy or electronic documents. The process may be manual (read and key) or electronic via a pattern recognition methodology.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Backup
Backup tape
DAT - digital audio tape
Digital audio tape
Disaster recovery tape
DLT - digital linear tape
Magnetic storage media
Media
QIC - quarter inch cartridge
Tape

Data Field

A name for an individual piece of standardized data to be extracted from an image collection. Fields can be the author of a document, a recipient, the date of a document or any other piece of data common to most documents in an image collection.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A unit of information in a database. Database records, for example, consist of an ordered list of fields where a specific kind of information is stored in each field. Fields are often printed as columns in database reports.

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Data Field Definition

Data field definition usually includes field structure (size of each field and whether it is a date, an integer, or a text field) and field organization (names and locations of data fields within a document record).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Data Format

The organization of information for display, storage, or printing. Data is maintained in certain common formats so that it can be used by various programs, which may only work with data in a particular format. This term is commonly used in the industry when asking another person about the state in which particular information exists. For example, "What format is it in, PDF or HTML?"

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

Data integrity

Refers to the validity of data. Data integrity can be compromised in a number of ways, including: human errors when data is entered, errors that occur when data is transmitted from one computer to another, software bugs or viruses, hardware malfunctions, such as disk crashes and natural disasters, such as fires and floods. There are many ways to minimize these threats to data including: backing up data on a regular basis, controlling access to data via security mechanisms, designing user interfaces that prevent the input of invalid data, and using error detection and correction software when transmitting data.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Data Lake

A storage repository holding large volumes of raw data. A warehouse for data of any size or type that is minimally processed and largely unanalyzed. Data lakes eliminate the up-front cost of processing data until those data are needed. Data lakes are intended to remove information

silos by combining data from multiple sources into a single repository and to avoid filtering, structuring, or otherwise "prejudging" the data before they can be fully analyzed.

Source: Herb Roitblat, Search 2020: The Glossary.

Data Mapping

Data mapping finds or suggests associations between files within a large body of data, which may not be apparent using other techniques.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Data Mining

"Data mining" generally refers to techniques for extracting summaries and reports from an organization's databases and data sets. In the context of electronic discovery, this term often refers to the processes used to cull through a collection of electronic data to extract evidence for production or presentation in an investigation or in litigation. Data mining can also play an important role in complying with data retention obligations under an organization's formal document management policies.

Source: Merrill Corporation, Electronic Discovery Glossary.

The process of extracting useful data from a volume of unstructured information. Data mining is used to search for patterns and systematic relationships in big data collections to extract other useful pieces of information from these collections.

Source: Herb Roitblat, Search 2020: The Glossary.

Data Protection Act (DPA)

This act implements a European Directive which among other things protects privacy and sets limits on what can be done with an individual's personal data. In particular, it places limitations on the transfer of such data between jurisdictions.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

Data Rate

The speed of a data communications channel, measured in bits per second.

Source: RSI, Glossary.

Data Set

Any group of files processed as a unit, and generally contained in one directory.

Source: Ibis Consulting, Glossary.

Data Stream

Data streams allow multiple forms of data to be associated with a file, including any number of graphic files, databases, programs, spreadsheets, word processing documents, or other data

types associated with a given file to alter some of the rules concerning computer security issues and computer forensics investigations.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Data Validation

A system for ensuring accuracy in data entry and consistency in formatting names and dates. Often accomplished by the use of validation tables to restrict entry of inconsistent or inaccurate data (e.g., date entered as 10/2/50, when it should be 01/02/50).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Database

A set of interrelated files stored electronically on a computer.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A collection of related data entered into individual records consisting of a number of different fields.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Information arranged in the computer in a rigorous, defined format to allow ease of recording and retrieval.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A collection of data arranged in tables along with reports, queries, and forms. Modern relational databases employ complex linkages among the data in the tables so that information can be entered only once, but still used to present coherent reports. A table is like a spread sheet, in which the columns correspond to fields and the rows to records, for example, individuals. Each field or column of the record indicates one piece of information about that individual.

See also:

Flat file database
Full text database
Relational database
SQL
WAIS - wide area information server

Database Administrator

A database administrator (short form DBA) is a person responsible for the installation, configuration, upgrade, administration, monitoring and maintenance of databases in an organization.

Source: http://en.wikipedia.org/wiki/Database_administrator.

Database Design

The process of deciding what database structure to use. Typically involves the construction of specific data fields and the overall design of bow the fields are to be used.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Database Management System (DBMS)

Software that controls the organization of a database and processes requests for database information from other applications.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Date Field

A data field in a database that contains the date of the document.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Date Filter

A filter option that allows for including/excluding specific dates or date ranges for application and/or mail users.

Source: Ibis Consulting, Glossary.

See also:

Extensions/sizes filter
Filter
MD5-known filter
Sender/recipient filter

Date Range Search

Date range search utilizes a document's metadata to find search results where the creation dates, access dates, or modification dates of documents fall within a specified range of dates. Refer to specific technology utilized to process ESI to determine the available dates based on file types and consider the handling of time zones during ESI processing.

Source: EDRM Search Glossary.

DBMS

See: Database Management System (DBMS)

DBX

Microsoft Outlook Express stores your messages in a folder that contains several different .dbx files. These files (folders.dbx, inbox.dbx, outbox.dbx) contain all your messages.

Source: Import messages into Windows Mail from Outlook Express, http://windows.microsoft.com/en-us/windows-vista/import-messages-into-windows-mail-from-outlook-express.

dd File

A "dd" file is a raw image file created using the dd forensic imaging tool, a command line program that uses command line arguments to control the imaging process.

Source: http://www.forensicswiki.org/wiki/Dd

De-Duplication

De-duplication ("de-duping") is the process of comparing electronic records based on their characteristics and removing duplicate records from the data set.

Source: Merrill Corporation, Electronic Discovery Glossary.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The process of providing one instance of an item when there was once two or more identical copies. This process usually involves landing all files into a database and then searching for duplicate files.

Source: RenewData, Glossary (10/5/2005).

The process of identifying (or some vendors includes actually removing) additional copies of identical documents in a document collection. There are three types of de-duplication: case, custodian, and production.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

The process of identifying (and/or removing) additional copies of identical documents in a document collection. There are three types of de-duplication: case, custodian, and production.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The method of data reduction that excludes duplicate messages (with their attachments) and files from further processing.

Source: Ibis Consulting, Glossary.

The process of removing duplicate records from a collection of data.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The process of determining which documents are duplicates. File systems can contain many copies of the same document, which need to be identified for efficiency's sake. Every time an email is sent it typically creates two additional copies of the email and its attachments, one in the sender's sent-items folder and once in the recipient's inbox. An email may also be sent to multiple recipients, thereby creating more copies.

See also:

Basic de-duplication	Duplicate	Horizontal Deduplication
Case de-duplication	Dynamic de-duplication	Production de-duplication
Custodian de-duplication	Global Deduplication	Vertical Deduplication

De-Shade

Remove shaded areas to render images more easily recognizable by OCR. De-shading software typically searches for areas with a regular pattern of tiny dots.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

De-Skew

A process where the computer detects and corrects the skew in an image file.

Source: RSI, Glossary.

The process of straightening skewed (off-center) images. De-skewing is one of the image enhancements that can improve OCR accuracy. Documents often become skewed when they are scanned or faxed.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Skew

De-Speckle

Remove isolated speckles from an image file. Speckles often develop when a document is scanned or faxed.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Decision Tree

A step-by-step method of distinguishing between Relevant and Non-Relevant Documents, depending on what combination of words (or other Features) they contain. A Decision Tree to identify Documents pertaining to financial derivatives might first determine whether or not a Document contained the word “swap.” If it did, the Decision Tree might then determine whether or not the Document contained “credit,” and so on. A Decision Tree may be created either through Knowledge Engineering or Machine Learning.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Decryption

Decryption is the process of converting encrypted data back into its original form, so it can be understood.

Source: Ibis Consulting, Glossary.

See also:

Encryption

Deduplication

A method of replacing multiple identical copies of a Document by a single instance of that Document. Deduplication can occur within the data of a single custodian (also referred to as Vertical Deduplication), or across all custodians (also referred to as Horizontal Deduplication).

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Deep Learning

An approach to building and training neural networks. Deep learning typically involves a hierarchical neural network where each of the levels in the hierarchy is trained separately.

Source: Herb Roitblat, Search 2020: The Glossary.

Default

A value or option assigned to data by a system when no specific value has been specified by an operator.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Deleted Data

Deleted data is data that, in the past, existed on the computer as live data and which has been deleted by the computer system or end-user activity. Deleted data remains on storage media in whole or in part until it is overwritten by ongoing usage or “wiped” with a software program specifically designed to remove deleted data. Even after the data itself has been wiped, directory entries, pointers, or other metadata relating to the deleted data may remain on the computer.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Deleted data are data that, in the past, existed on the computer as live data and which have been deleted by the computer system or end-user activity. Deleted data remain on storage media in whole or in part until they are overwritten or “wiped.” Even after the data itself have been wiped, directory entries, pointers or other metadata relating to the deleted data may remain on the computer.

Source: Merrill Corporation, Electronic Discovery Glossary.

Recoverable information from deleted files and data may be stored in unallocated or slack space on a computer hard drive.

Source: RenewData, Glossary (10/5/2005).

Data that at one time existed on a computer system as live data but that has been deleted, however, such deleted data inhabits storage media in some form until it is overwritten or "wiped" with a software program specifically designed to remove deleted data. Once the storage media has been wiped, directory entries, pointers, or other metadata often remain.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Data that once existed on a computer and has subsequently been deleted by the user. Deleted data actually remains on the computer until it is overwritten by new data or “wiped” with a specific software program. (Even after wiping, metadata such as directory entries or pointers may still remain.)

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Deletion

Deletion is the process whereby data is removed from active files and other data storage structures on computers and rendered inaccessible except using special data recovery tools

designed to recover deleted data. Deletion occurs in several levels on modern computer systems:

1. File level deletion: Deletion on the file level renders the file inaccessible to the operating system and normal application programs and marks the space occupied by the file's directory entry and contents as free space, available to reuse for data storage.
2. Record level deletion: Deletion on the record level occurs when a data structure, like a database table, contains multiple records; deletion at this level renders the record inaccessible to the database management system (DBMS) and usually marks the space occupied by the record as available for reuse by the DBMS, although in some cases the space is never reused until the database is compacted. Record level deletion is also characteristic of many e-mail systems.
3. Byte level deletion: Deletion at the byte level occurs when text or other information is deleted from the file content (such as the deletion of text from a word processing file); such deletion may render the deleted data inaccessible to the application intended to be used in processing the file, but may not actually remove the data from the file's content until a process such as compaction or rewriting of the file causes the deleted data to be overwritten.

   Source: Merrill Corporation, Electronic Discovery Glossary.

Deletion is the process whereby data is removed from active files and other data storage structures on computers and rendered inaccessible except using special data recovery tools designed to recover deleted data.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Removing active files making them unavailable. Special data recovery tools can still retrieve these files.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Descender

The portion of a character which falls below the main part of the letter (e.g. g, p,q).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Desktop

Usually refers to an individual PC -- a user's desktop computer.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Desktop Publishing

PC systems used to prepare direct print output or output suitable for printing presses.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Destination File

The file that a linked or embedded object is inserted into, or that data is saved to. The source file contains the information that is used to create the object. When you change information in a destination file, the information is not updated in the source file.

Source: Glosbe, destination file, https://en.glosbe.com/en/en/destination%20file

DIA/DCA (Document Interchange Architecture)

An IBM standard for transmission and storage of voice, text or video over networks.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Diacritic Specification

A diacritic specification is a phonetic marker added to letter (above or below) indicating a change in the way it is to be pronounced or stressed. For languages that include diacritic characters on certain characters (such as vowels), specifying whether the diacritics should match is a search option.

Source: EDRM Search Glossary.

Digital Audio Tape

See: DAT (Digital Audio Tape)

Digital Linear Tape (DLT)

Digital linear tape is a form of magnetic tape and drive system used for computer data storage and archiving. DLT is one of several technologies developed in recent years to increase the data-transfer rates and storage capacities of computer tape drives.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing SearchStorage.com, http://searchstorage.techtarget.com/sDefinition/0,,sid5_gci759350,00.html.

A document storage medium in a cartridge. DLT tapes are often used for backup tapes.

See also:

Backup
Backup tape
DAT - digital audio tape
Data extraction
Digital audio tape
Disaster recovery tape
Magnetic storage media
Media
QIC - quarter inch cartridge
Tape

Digital Signal Processor (Processing) (DSP)

A special purpose computer (or technique) which digitally processes signals and electrical/analog waveforms.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Digital to Analog Converter

See: DAC (Digital to Analog Converter)

Digital Versatile Disc (DVD)

A plastic disc, like a CD, on which data can be written and read. DVDs are faster, can hold more information, and can support more data formats than CDs.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

An optically encoded removable disc similar to a CD, but with much higher capacity.

See also:

CD	DVD-ROM	Magnetic storage media
CD-R	Floppy disk	Media
CD-ROM	Hard disk	Optical disk
CD-RW	Hard drive	Storage media
Disc	Jaz disk	WORM disk
Disk	Laser disc	Zip disk
Diskette	Magnetic disk	

Digital Video Disc (DVD)

See: Digital Versatile Disc (DVD)

Dimensionality Reduction

A Feature Engineering method used to reduce the total number of Features considered by a Machine Learning Algorithm. Simple Dimensionality Reduction methods include Stemming and Stop Word elimination. More complex Dimensionality Reduction methods include Latent Semantic Indexing and Hashing.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Directory

A directory is, in general, an approach to organizing information, the most familiar example being a telephone directory.

Source: TechTarget, directory definition, http://searchwindowsserver.techtarget.com/definition/directory

Dirty OCR

Electronic documents resulting from inaccurate optical character recognitions.

See also:

ICR

OCR

Optical Character Recognition

Pattern recognition

Disaster Recovery Plan

A plan devised by an organization to avoid data loss or business disruption following a power loss, a natural disaster, or an act of terrorism. A good disaster recovery plan calls for daily backup functions and contingency plans to minimize data loss and business interruption.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Disaster Recovery Tape

Portable media used to store data that is not presently in use by an organization to free up space but still allow for disaster recovery. May also be called "backup tapes."

Source: Merrill Corporation, Electronic Discovery Glossary.

See also:

Backup

Backup tape

DAT - digital audio tape

Data extraction

Digital audio tape

DLT - digital linear tape

Magnetic storage media

Media

QIC - quarter inch cartridge

Tape

Disc

An optical disc.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Round, flat storage media with layers of material which enable the recording of data.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

CD
CD-R
CD-ROM
CD-RW
Disk
Diskette
DVD
DVD-ROM
Floppy disk
Hard disk
Hard drive
Jaz disk
Laser disc
Magnetic disk
Magnetic storage media
Media
Optical disk
Storage media
WORM disk
Zip disk

Disclosure

Disclosure means the giving out of information, either voluntarily or to be in compliance with legal regulations or workplace rules.

Source: EDRM Presentation Guide.

Discovery

A pre-trial process in which each party tries to find all the information held by the other party and by certain third parties that is relevant, probative and can be admitted into evidence at trial. Each party is required to cooperate with the other to the extent required by the relevant rules of civil procedure.

Source: RenewData, Glossary (10/5/2005).

The pre-trial procedure by which each party gains information held by the adverse party concerning a case. Discovery is also the disclosure of facts, documents, electronically stored information and tangible objects by an adverse party.

Source: Ibis Consulting, Glossary.

The process of providing documents to an opposing party in US litigation. Unlike the UK, the US process is driven by "pull" i.e. a party needs to specify the documents they want each opposing party to produce.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

See also:

Computer evidence
Computer forensics
Computer investigations
Electronic discovery / e-discovery
Electronic evidence
Forensic analysis
Forensics
Mirroring

Discovery Request

An official request, by the opposing attorney, to deliver documents relevant to particular case issues.

See also:

Document request
Interrogatory
Request for admission

Discovery Tracking

The use of a database to monitor the progress of discovery as well as the content and consistency of discovery responses.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Disk

A storage medium capable of storing large amounts of data. Disk types include magnetic disks (both hard disks and floppy disks) and optical disks.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A magnetic floppy or hard disk.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Round, flat storage media with layers of material which enable the recording of data.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

It may be a floppy disk, or it may be a hard disk. Either way, it is a magnetic storage medium on which data is digitally stored. A disc may also refer to a CD-ROM.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

CD
CD-R
CD-ROM
CD-RW
Disc
Diskette
DVD
DVD-ROM
Floppy disk
Hard disk
Hard drive
Jaz disk
Laser disc
Magnetic disk
Magnetic storage media
Media
Optical disk
Storage media
WORM disk
Zip disk

Disk Drive

The device that houses a disk and controls the connection between the computer and the magnetic disk.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Floppy disk drive	Portable drive	Zip drive
Jaz drive	Storage device	
Magneto-optical drive	Tape drive	

Disk Mirroring

When files are stored on a computer system's hard disk, a "mirror" copy is made on an additional hard disk or a separate part of the same disk to safeguard information in case of disaster.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A method of data backup that copies or “mirrors” each saved file on a hard disk onto a second hard disk.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Disk Operating System (DOS)

Acronym for disk operating system. The term DOS can refer to any operating system, but it is most often used as a shorthand for MS-DOS (Microsoft disk operating system). Originally developed by Microsoft for IBM, MS-DOS was the standard operating system for IBM-compatible personal computers.

Source: http://www.webopedia.com/TERM/D/DOS.html

A set of programs that controls the computer and supports software applications. MS-DOS (Microsoft Disk Operating System) is popular because it was the system used in the original IBM PC and subsequent clones.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Linux	Network operating system	OS
Microsoft DOS	NOS	UNIX
Microsoft Windows	Operating system	Windows

Xenix

Diskette

Synonym for “floppy disk.”

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

CD	DVD-ROM	Magnetic storage media
CD-R	Floppy disk	Media
CD-ROM	Hard disk	Optical disk
CD-RW	Hard drive	Storage media
Disc	Jaz disk	WORM disk
Disk	Laser disc	Zip disk
DVD	Magnetic disk	

Distributed Data

Data that resides on portable media and non-local devices such as laptop computers, home computers, CD-ROMs, floppy disks, zip drives, wireless communication devices, personal digital assistants (PDAs), web pages, Internet repositories such as email hosted by Internet service providers or portals, and the like that belongs to the organization and not the user.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Information which resides on non-local devices such as home computers, laptop computers, PDAs, or even Internet repositories.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

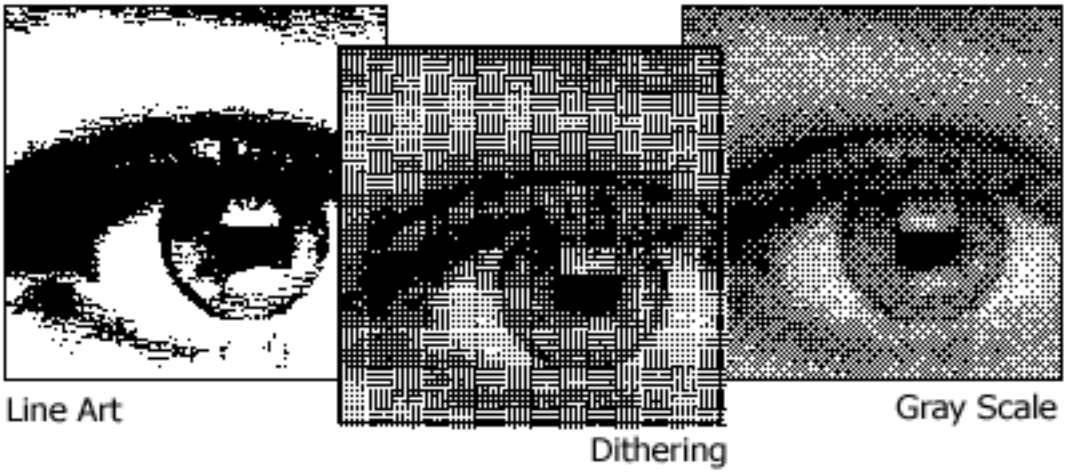
Distributed data is that information belonging to an organization which resides on portable media and non-local devices such as home computers, laptop computers, floppy disks, CD-ROMs, personal digital assistants (“PDAs”), wireless communication devices (i.e., Blackberry), zip drives, Internet repositories such as e-mail hosted by Internet service providers or portals, Web pages, and the like. Distributed data also includes data held by third parties such as application service providers and business partners.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Dithering

Creating the illusion of new colors and shades by varying the pattern of dots. Newspaper photographs, for example, are dithered. If you look closely (see example below), you can see

that different shades of gray are produced by varying the patterns of black and white dots. There are no gray dots at all. The more dither patterns that a device or program supports, the more shades of gray it can represent. In printing, dithering is usually called halftoning, and shades of gray are called halftones.



Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

DLT

See: Digital Linear Tape (DLT)

DMS (Document Management System)

Essentially a database to store and retrieve firm documents by client/matter number, title, author, date and/or keywords.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Document

Any file produced by a software application.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Any file produced by a software application. Includes but is not limited to any electronically stored data on magnetic or optical storage media as an "active" file or files (readily readable by one or more computer applications or forensics software); any "deleted" but recoverable electronic files on said media; any electronic file fragments (files that have been deleted and partially overwritten with new data); and slack (data fragments stored randomly from random access memory on a hard drive during the normal operation of a computer [RAM slack] or residual data left on the hard drive after new data has overwritten some but not all of previously stored data).

Source: RSI, Glossary.

See Rule 34 of the Federal Rules of Civil Procedure.

Source: Merrill Corporation, Electronic Discovery Glossary.

One or several single pages of images that make a logical single communication of information. Examples include a letter, a report, a memo or an airline ticket. A "document" may be any means of communicating, informing or educating, including hard-copy paper, electronic documents, e-mail, voice mail, video, x-rays, drawings, etc.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A document is a page or collection of pages that are physically or logically (or both) linked.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Fed. R. Civ. P. 34(a) defines a document as "including writings, drawings, graphs, charts, photographs, phonorecords, and other data compilations." In the electronic discovery world, a document also refers to a collection of pages representing an electronic file. E-mails, attachments, databases, word documents, spreadsheets, and graphic files are all examples of electronic documents.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

In the context of Electronic Discovery, a discrete item of Electronically Stored Information that may be the subject or result of a search or review effort.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Document Assembly

Software function that gathers facts about a client, then merges data and text to draft a unique document for that client that varies depending on the facts of each case. Typically is performed by answering a series of questions or extracting data from a database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Document Boundaries

The beginning and ending pages of a document. A folder full of papers may have no obvious indication where one document ends and another begins (e.g., if the staples or paperclips that held the documents together were removed). When paper documents are scanned, the boundaries between the documents usually has to be determined and noted.

Document Collection

The process of gathering Electronically Stored Information for search, review, and production; the set of Documents resulting from such a process. In many cases, the Document Collection and Document Population are the same; however, it is important to note that Document Population refers to the set of Documents over which a particular Statistical Estimate is calculated, which may be the entire Document Collection, a subset of the Document Collection (e.g., the documents with a particular file type or matching particular Search Terms), a superset of the Document Collection (e.g., the universe from which the Document Collection was gathered), or any combination thereof.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Document Date

The original creation date of a document usually noted on the document itself. In the case of a letter, when the letter was written indicated by the date of the letter. On an email indicated by the date-stamp of the email.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Document Depository

A library of hard copies of all documents in a specific case, sometimes the originals, and often run under guidelines specified by the court.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A central library of all documents in a case, either hard copies or images, with some form of electronic access.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Document Enhancement

A context-sensitive annotation to a full-text document.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Document Interchange Architecture

See: DIA/DCA (Document Interchange Architecture)

Document Management System

See: DMS (Document Management System)

Document Metadata

Data stored in a document about the document. Often this data is not immediately viewable in software application used to create/edit the document, but often can be accessed via a "properties" view. Contrast with file system metadata and email metadata.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

Data stored in a document about the document. Often this data is not immediately viewable in software application used to create/edit the document, but often can be accessed via a "properties" view. Example: Last Accessed Date, Last Edited By, Users, etc.

See also:

Customer-added metadata	File parameters	General metadata
Email metadata	File system metadata	Metadata
Extrinsic data	File-specific metadata	Vendor-added metadata

Document Number

Similar to a Bates number, a document number is a unique identifier assigned to a document or file. Document numbers are used to track documents or files throughout one or more lawsuits or similar proceedings.

See also:

Bates number	Bates stamp
Bates prefix	Bates stamping

Document Population

The set of Electronically Stored Information or Documents about which a Statistical Estimation may be made.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Document Request

In a lawsuit or similar proceeding, a written request from a part to another party or to a non-party asking the recipient to produce or permit the inspection of documents, electronically stored information, or tangible objects.

External links:

Rule 34. Producing Documents, Electronically Stored Information, and Tangible Things, or Entering onto Land, for Inspection and Other Purposes, https://www.law.cornell.edu/rules/frcp/rule_34

See also:

Discovery Request Interrogatory Request for admission

Document Retention Policy

A set of rules for determining how long documents of different types need to be retained for business or legal purposes. These rules vary by business type, document type, and counsel strategies. A document retention policy is one part of a records management scheme.

Document Segments

Documents may be split into multiple segments (such as Abstract, Body, Title, References, Citation, etc.). The Boolean operators may be limited to a specific document segment. In these situations, you may need to specify the search scope of the document.

Source: EDRM Search Glossary.

Document Sizes

(U.S.):

- A Size 8.5" by 11" (A4)
- B Size 11" by 17" (A3)
- C Size 17" by 22" (A2)
- D Size 24" by 36" (A1)
- E Size 36" by 48" (A0)

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Document Template

Sets of index fields for documents.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Document Type

A typical field used in bibliographical coding. Typical document type examples include letter, memo, report, article and others. Often referred to as Doc Type.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

DOS

See: Disk Operating System (DOS)

DOS Prompt

Usually a disk drive letter followed by the greater than (>) symbol. It is the position from which DOS functions are executed manually if the computer has no common user interface (CUI) or Graphical User Interface (GUI).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Dot Pitch

Distance of one pixel in a CRT to the next pixel on the vertical plane. The smaller the number, the higher quality display.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Dots Per Inch (DPI)

A measurement of scanner resolution. The number of pixels a scanner can physically distinguish in each vertical and horizontal inch of an original image. Documents are normally scanned at a resolution of between 200 dpi and 400 dpi.

SOURCE:RSI, Glossary.

Double-Sided Scanner

Double-sided scanning uses a single-sided scanner to scan double-sided pages, scanning one collated stack of paper, then flipping it over and scanning the other side.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Duplex scanner
Flatbed scanner
Scanner
Simplex scanner

Download

To transfer data to the user's computer from another computer.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

DPA

See: Data Protection Act (DPA)

DPI

See: Dots Per Inch (DPI)

Drag-and-Drop

A common way to move or copy a file or folder is to highlight it and literally "drag" a copied version of it to another location. First the mouse would be used to highlight the file. Then while holding down the left mouse button, the name of the file would be dragged to a new location. In the background, the operating system creates a new copy and places it in the new location. For example, you can drag a file to the Recycle Bin to delete the file, or to a folder to copy or move it to that location.

Source: EDRM Collection Standards

The movement of on-screen objects by dragging them across the screen with the mouse.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

DRAM (Dynamic Random Access Memory)

A memory technology which is periodically "refreshed" or updated – as opposed to "static" RAM chips which do not require refreshing. The term is often used to refer to the memory chips themselves. Varieties are:

- CDRAM: Cache DRAM (contains static cache)
- EDODRAM: Extended data out DRAM
- EDRAM: Enhanced DRAM (contains a static memory buffer and cache controller)
- SDRAM: Synchronous DRAM (added clock and burst addressing capability)
- SGRAM: Synchronous Graphics RAM (a single port SDRAM)
- WRAM: Window RAM (dual port video RAM)
- VRAM: Video RAM (a dual ported DRAM, good for graphics)

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Memory RAM ROM

Dropped Items

Another form of validation utilized to ensure that Responsive items are not being inadvertently omitted through changes to the search criteria. As the search criteria set is being updated and modified during the initial investigation and analysis, a comparison would sample documents that were originally results of one search criteria set but are no longer results of the modified search criteria set. If Responsive documents are found upon review of dropped items, special attention should be paid to determine whether additional terms need to be created to capture these items or if modifications made to the criteria should be changed so these or similar items would be included in the results.

Source: EDRM Search Glossary.

DSP

See: Digital Signal Processor (Processing) (DSP)

Duces Tecum

Latin for "bring with thee." Usually means come to a deposition or court appearance with documents.

Source: Ibis Consulting, Glossary.

Dumb Terminal

A network terminal with keyboard, monitor, and network interface but no hard drive for independent processing capability.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Duplex

The ability of a scanner to scan both sides of a sheet simultaneously. Requires two scanner cameras and often two processing boards.

Source: RSI, Glossary.

Two-sided page(s).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Duplex Scanner

Duplex scanners automatically scan both sides of a double-sided page, producing two images at once.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Double-sided scanner
Scanner
Flatbed scanner
Simplex scanner

Duplicate

An exact duplicate of another document in a database. Duplicates typically arise when multiple document productions from separate sources are coded and contain copies of the same documents.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Basic de-duplication
Case de-duplication
Custodian de-duplication
De-duplication
Dynamic de-duplication
Global Deduplication
Horizontal Deduplication
Production de-duplication
Vertical Deduplication

Duty

Legal obligation for managing the "Risk" associated with specific information. Legal and RIM have the responsibility for legal duties and obligations (i.e., legal hold preservation and regulatory retention obligations).

Source: IGRM White Paper

DVD

See: Digital Versatile Disc (DVD)

DVD-ROM

Digital versatile disc-read only memory (DVD-ROM) is a read-only digital versatile disc (DVD) commonly used for storing large software applications. It is similar to a compact disk-read only memory (CD-ROM) but has a larger capacity. A DVD-ROM stores around 4.38 GB of data. A CD-ROM usually stores 650 MB of data.

Source: Technopedia, Digital Versatile Disc-Read Only Memory (DVD-ROM), https://www.techopedia.com/definition/24480/digital-versatile-disc-read-only-memory--dvd-rom

See also:

CD
CD-R
CD-ROM

CD-RW	Hard disk	Media
Disc	Hard drive	Optical disk
Disk	Jaz disk	Storage media
Diskette	Laser disc	WORM disk
DVD	Magnetic disk	Zip disk
Floppy disk	Magnetic storage media	

Dynamic De-Duplication

A proprietary dynamic de-duplication technology designed to handle large volumes of information and ensure that unique meta-data and original content are stored only once. This dynamic de-duplication process is executed as the data flows off the tapes, in order to avoid large and expensive processing and storage requirements.

Source: RenewData, Glossary (10/5/2005).

See also:

Basic de-duplication	Custodian de-duplication	Duplicate
Case de-duplication	De-duplication	Production de-duplication

Dynamic Random Access Memory

See: DRAM (Dynamic Random Access Memory)

E

E-Mail (Electronic Mail)

Electronic mail, commonly referred to as "e-mail" or "email," is an electronic means for communicating information under specified conditions, generally in the form of text messages, through systems that will send, store, process, and receive information and in which messages are held in storage until the addressee accesses them.

Source: Merrill Corporation, Electronic Discovery Glossary.

A simple text message -- a piece of text sent to a recipient. In the beginning and even today, e-mail messages tend to be short pieces of text, although the ability to add attachments now makes many e-mail messages quite long. Even with attachments, however, e-mail messages continue to be text messages.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing How Stuff Works, http://computer.howstuffworks.com/email.htm/printable.

The whole of an electronic document containing the message envelope and message content (attachments, etc.).

Source: RenewData, Glossary (10/5/2005).

Electronic mail, or computer-based mail.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

Any data set or group of files originating from mail containers or e-mail systems. This includes single-mail items outside of their mail applications, like MSG (Outlook) and EML files (RFC822 single mail containers), RFC822 mail folders as well as multi-mail archives (PST, NSF, etc.).

Source: Ibis Consulting, Glossary.

Early Case Assessment (ECA)

An industry-specific term generally used to describe a variety of tools or methods for investigating and quickly learning about a Document Collection for the purposes of estimating the risk(s) and cost(s) of pursuing a particular legal course of action.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A widely abused term in which corporate data is sifted and categorised with a view to determining an organisation's exposure in the context of a dispute. The best ECA systems allow the sifting to take place within a corporation's own data store and can be used to drill down rapidly to identify the most pertinent evidentiary material and to facilitate decisions whether to litigate or settle.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

EB (Exabyte)

1 million terabytes. The US business community is estimated to have created 35-50 exabytes of electronic data in 2004.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Bit
Byte
KB - kilobyte
MB - megabyte
GB - gigabyte
TB - terabyte
PB - petabyte

EDI-Oracle Study

An ongoing initiative (as of January 2013) of the Electronic Discovery Institute to evaluate participating vendors' search and document review efforts using a Document Collection contributed by Oracle America, Inc.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

EDMS (Electronic Document Management System)

EDMS - electronic document management system is a software program that manages the creation, storage and control of documents electronically. The primary function of an EDMS is to manage electronic information within an organization workflow. A basic EDMS should include document management, workflow, text retrieval, and imaging. An EDMS must be capable of providing secure access, maintaining the context, and executing disposition instructions for all records in the system.

Source: EDMS - Electronic Document Management System, http://www.edms.net

EDRM

EDRM is an organization that creates practical resources to improve e-discovery and information governance. Since 2005 the e-discovery community has relied on EDRM for leadership, standards, best practices, tools, guides and test data sets to improve electronic discovery and information governance. Member individuals, law firms, corporations and government organizations actively contribute to the direction of EDRM.

EDRM Diagram

The EDRM diagram is conceptual, non-linear, iterative model of the e-discovery process.

It represents a conceptual view of the e-discovery process, not a literal, linear or waterfall model. One may engage in some but not all of the steps outlined in the diagram, or one may elect to carry out the steps in a different order than shown here.

The diagram also portrays an iterative process. One might repeat the same step numerous times, honing in on a more precise set of results. One might also cycle back to earlier steps, refining one's approach as a better understanding of the data emerges or as the nature of the matter changes.

Source: EDRM Diagram Elements

EDRM Framework

See: EDRM Diagram

EDRM Framework Guides

The EDRM framework guides are a series of practical guides developed for each stage of the e-discovery process as depicted in the EDRM framework.

EDRM Phases

The Electronic Discovery Reference Model, also referred to as EDRM or the EDRM diagram, outlines the key processes and stages of the e-discovery process in the form of nine interrelated phases: Information Governance, Identification, Preservation, Collection, Processing, Review, Analysis, Production, and Presentation. Each phase represents a core stage of the e-discovery process. By breaking the e-discovery process into phases, practitioners can leverage core resources (i.e. people, technology, and processes) in a more organized fashion to achieve desired results.

Source: EDRM Metrics Glossary

EGA (Enhanced Graphics Adapter)

Short for Enhanced Graphics Adapter, EGA is a video standard manufactured by IBM in 1984 with a higher resolution (640 x 350) and more colors (16 from a palette of 64) when compared to earlier standards such as CGA.

Source: Computer Hope, EGA definition, http://www.computerhope.com/jargon/e/ega.htm

The Enhanced Graphics Adapter (EGA) is a historical IBM PC computer display standard from 1984 that superseded and exceeded the capabilities of the CGA standard introduced with the original IBM PC, and was itself superseded by the VGA standard in 1987.

Source: Wikipedia, Enhanced Graphics Adapter, https://en.wikipedia.org/wiki/Enhanced_Graphics_Adapter

EIA (Electronic Industries Association)

A trade association.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

EISA (Extended Industry Standard Architecture)

One of the standard buses used for PCs.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Electronic Discovery / E-Discovery

Discovery documents produced in electronic formats rather than hardcopy. The production may be contained on hard drives, tapes, CDs, DVDs, external hard drives, etc. Once received,

these documents are converted to .tif format. It is during the conversion process that metadata can be extracted.

Source: RSI, Glossary.

A process that includes electronic documents and email into a collection of "discoverable" documents for litigation. Usually involves both software and a process that searches and indexes files on hard drives or other electronic media. Extracts metadata automatically for use as an index. May include conversion of electronic documents to an image format as if the document had been printed out and then scanned.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

The discovery of electronic documents and data including e-mail, Web pages, word processing files, computer databases, and virtually anything that is stored on a computer. Technically, documents and data are “electronic” if they exist in a medium that can only be read through the use of computers. Such media include cache memory, magnetic disks (such as computer hard drives or floppy disks), optical disks (such as DVDs or CDs), and magnetic tapes.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The process of finding, identifying, locating, retrieving, and reviewing potentially relevant data in designated computer systems.

The process of identifying, preserving, collecting, processing, searching, reviewing and producing Electronically Stored Information that may be Relevant to a civil, criminal, or regulatory matter.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

See also:

Computer evidence	Discovery	Forensics
Computer forensics	Electronic evidence	Mirroring
Computer investigations	Forensic analysis	

Electronic Document

A document that has been scanned, or was originally created on a computer. Documents become more useful when stored electronically because they can be widely distributed instantly, and allow searching. HTML and PDF are well known electronic document formats.

Source: RSI, Glossary.

Electronic Document Discovery

See: Electronic Discovery / E-Discovery

Electronic Document Management System

See: EDMS (Electronic Document Management System)

Electronic Evidence

According to Black's law dictionary, evidence is "any species of proof, or probative matter, legally presented at the trial of an issue, by the act of the parties and through the medium of witnesses, records, documents, exhibits, concrete objects, etc. for the purpose of inducing belief in the minds of the court or jury as to their contention." Electronic information (like paper) generally is admissible into evidence in a legal proceeding.

Source: RenewData, Glossary (10/5/2005).

Any computer-generated data that is relevant to a case. Included are email, text documents, spreadsheets, images, database files, deleted email and files and back-ups. The data may be on floppy disk, zip disk, hard drive, tape, CD or DVD.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Norcross Group FAQ's, http://norcrossgroup.com/faq.html#5.

See also:

Computer evidence	Discovery	Forensic analysis
Computer forensics	Electronic discovery / e-discovery	Forensics
Computer investigations		Mirroring

Electronic Industries Association

See: EIA (Electronic Industries Association)

Electronic Mail

See: E-Mail (Electronic Mail)

Electronic Mail Message

Commonly referred to as "e-mail", an electronic mail message is a document created or received via an electronic mail system, including brief notes, formal or substantive narrative documents, and any attachments, such as word processing and other electronic documents, which may be transmitted with the message.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Electronic Record

Information recorded in a form that requires a computer or other machine to process it and that otherwise satisfies the definition of a record.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Elusion

The fraction of Documents identified as Non-Relevant by a search or review effort that are in fact Relevant. Elusion is estimated by taking a Random Sample from the Null Set and determining how many (or what Proportion of) Documents are actually Relevant. A low Elusion value has commonly been advanced as evidence of an effective search or review effort (see, e.g., Kleen), but that can be misleading because it quantifies only those Relevant Documents that have been missed by the search or review effort; it does not quantify the Relevant Documents found by the search or review effort (i.e., Recall). Consider, for example, a Document Population containing one million Documents, of which ten thousand (or 1%) are Relevant. A search or review effort that returned 1,000 Documents, none of which were Relevant, would have 1.001% Elusion, belying the failure of the search. Elusion = 100% – Negative Predictive Value.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

An information retrieval measure of the proportion of responsive documents that have been missed. Most often used as a quality assurance measure in which a sample of non-retrieved documents is evaluated to determine whether a review has met reasonable criteria for completeness.

Source: Herb Roitblat, Predictive Coding Glossary.

Em

In any print font or size is equal to the width of the letter "M" in that font and size.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Email

See: E-Mail (Electronic Mail)

Email Address

An electronic mail address. Email addresses follow the formula: user-ID@domain-name. In some email systems, a user's email address is "aliased" or represented by their natural name rather than their fully qualified email address. For example, john.doe@abc.com might appear simply as John Doe.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

Email Attachment

Electronic files that are sent along with an email.

Email Message Store

A top most e-mail message store is the location in which an e-mail system stores its data. For instance, an Outlook PST (personal storage folder) is a type of top most file that is created when a user's Microsoft Outlook mail account is set up. Additional Outlook PST files for that user can be created for backing up and archiving Outlook folders, messages, forms and files. Similar to a filing cabinet, which is not considered part of the paper documents contained in it, a top most store generally is not considered part of a family.

Source" Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Email Metadata

Data stored in an email about the email. Often this data is not even viewable in email client application used to create the email. The amount of email metadata available for a particular email varies greatly depending on the email system. Contrast with file system metadata and document metadata.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary

Data stored in an email about the email. Often this data is not even viewable in email client application used to create the email. The amount of email metadata available for a particular email varies greatly depending on the email system. Example: Subject, Sent Date, To, Attachment, etc.

See also:

Customer-added metadata	File parameters	General metadata
Document metadata	File system metadata	Metadata
Extrinsic data	File-specific metadata	Vendor-added metadata

Email Threading

Grouping together email messages that are part of the same discourse, so that they may be understood, reviewed, and coded consistently as a unit.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Embed

To insert an object in its native format into a compound document.

Source: Ibis Consulting, Glossary.

Embedded Chart

A chart or a graph that would normally be displayed within a spreadsheet, but that may block underlying text or data.

Embedded Object

A data container without file properties allocated within another file such as a graphic, MS Word, or MS Excel file (having the properties of that object), typically represented by an application-specific icon in the body of any TIFF'ed content. This includes information that is contained in a source file and inserted into a destination file. Once embedded, the object becomes a part of the destination file.

Source: Ibis Consulting, Glossary.

See also:

Bibliographic coding	Link source	Object
Link object	Linked object	

EML

EML is a file extension for an e-mail message saved to a file in the MIME RFC 822 standard format by Microsoft Outlook Express as well as some other email programs.

Source: EML File Format, http://whatis.techtarget.com/fileformat/EML-Microsoft-Outlook-Express-mail-message-MIME-RFC-822.

A single RFC822 mail file message.

Source: Ibis Consulting, Glossary

An email file format, usually containing a single email message.

See also:

Container	NSF	Single-mail archive
Mail container	OST	Single-mail container
Mailbox	PST	SMTP
MSG	RFC compliant email	
Multi-mail container	RFC822	

Empty File

A file with no content, but with file properties, structure, and typically also metadata (including commercial application data).

Source: Ibis Consulting, Glossary.

Emulate

To imitate a device with a second device using a graphical user interface.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

En

Half the width of an Em.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Encapsulated PostScript (EPS)

Uncompressed files for images, text and objects. Only print on PostScript printers.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Encryption

A technology that renders the contents of a file unintelligible to anyone not authorized to read it. Encryption is used to protect information as it moves from one computer to another.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A technology that renders the contents of a file unintelligible to anyone not authorized to read it. Encryption is used to protect information as it moves from one computer to another, and is an increasingly common way of sending credit card numbers over the Internet when conducting e-commerce transactions.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

The conversion of data into a form called ciphertext that cannot be easily understood by unauthorized people/applications.

Source: Ibis Consulting, Glossary.

The coding of messages to increase security and make transmission only readable by recipients with the ability to decode only by using the same algorithms.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A procedure that renders the contents of a message or file unintelligible to anyone not authorized to read it.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Decryption

End Document Number

The last single page image of a document. Often called End Doc#.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

End of File (EOF)

A distinctive code which uniquely marks the end of a data file.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

End User Program

The program used to perform searches, viewing and retrieval of a scanned and/or coded collection of images. Examples include Summation, Concordance, JFS Litigators Notebook, Ringtail, Paradox, InMagic DB/Textworks and many others.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Endorser

A little printer in a scanner that adds a document-control number to each scanned sheet. Some forms control processing software can control this printer.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Enhanced Graphics Adapter

See: EGA (Enhanced Graphics Adapter)

Enhanced Parallel Port (EPP)

Also known as Fast Mode Parallel Port. A new, industry standard parallel port, having high transfer times competitive with SCSI.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Enhanced Small Device Interface (ESDI)

A defined, common electronic interface for transferring data between computers and peripherals, particularly disk drives.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

EOF

See: End of File (EOF)

EORHB

EORHB v. HOA Holdings LLC, Civ. Action No. 7409-VCL, tr. and slip op. (Del. Ch. Oct. 19, 2012). The first case in which a court *sua sponte* directed the parties to use Predictive Coding as a replacement for Manual Review (or to show cause why this was not an appropriate case for Predictive Coding), absent either party's request employ Predictive Coding. Vice Chancellor J. Travis Laster also ordered the parties to use the same E-Discovery vendor and to share a Document repository.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

EPP

See: Enhanced Parallel Port (EPP)

EPS

See: Encapsulated PostScript (EPS)

eRecall

An estimate of Recall computed from prevalence and production frequency.

Source: Herb Roitblat, Predictive Coding Glossary.

Error / Error Rate

The fraction of all Documents that are incorrectly coded by a search or review effort. Note that Accuracy + Error = 100%, and that 100% – Accuracy = Error. While a low Error Rate is commonly advanced as evidence of an effective search or review effort, its use can be misleading because it is heavily influenced by Prevalence. Consider, for example, a Document Population containing one million Documents, of which ten thousand (or 1%) are Relevant. A search or review effort that found none of the relevant Documents would have 1% Error, belying the failure of the search or review effort.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Escaping Mechanism

To search a keyword which contains a wildcard character such as a question mark, an escaping mechanism is needed to search. Availability of multi-character wildcards may be limited in some systems. Some search engines require a certain number of leading characters and do not support search terms that start with a wildcard.

Source: EDRM Search Glossary.

ESDI

See: Enhanced Small Device Interface (ESDI)

ESI / Electronically Stored Information

Electronically Stored Information or ESI is information that is stored electronically on enumerable types of media regardless of the original format in which it was created.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Electronically Stored Information: this is an all inclusive term referring to conventional electronic documents (e.g. spreadsheets and word processing documents) and in addition the contents of databases, mobile phone messages, digital recordings (e.g. of voicemail) and transcripts of instant messages. All of this material needs to be considered for disclosure.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

Used in Federal Rule of Civil Procedure 34(a)(1)(A) to refer to discoverable information “stored in any medium from which the information can be obtained either directly or, if necessary, after translation by the responding party into a reasonably usable form.” Although Rule 34(a)(1)(A) references “Documents or Electronically Stored Information,” individual units of review and production are commonly referred to as Documents, regardless of the medium.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Ethernet

A common way of networking PCs to create a LAN.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Evaluation Order

While the evaluation order should be immaterial, some search engines produce different results if the order is specified differently. In other implementations, the performance of search is impacted by the order of specification.

Source: EDRM Search Glossary.

Exabyte

See: EB (Exabyte)

Exceptions Report

A report listing documents that could not be processed within the parameters of the normal electronic discovery processing. For various reasons, the documents listed in the exceptions report could not be opened, their text extracted, or they could not be properly imaged. Effective systems minimize exceptions, because these documents may require special processing making them more expensive and delaying the production process.

Expansion Slot

A space inside a computer used to connect a board that controls other functions, such as a scanner or modem, to the motherboard.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Experimental Design

A standard procedure accepted in the scientific community for the evaluation of competing hypotheses. There are many valid experimental designs. Some that can be appropriate for evaluating Technology-Assisted Review processes include Crossover Trials and Parallel Trials.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Exploratory Search

Ad hoc or single logical query, likely to be employed in knowledge management effort on the left side, or as ad hoc search as part of case assessment, review or post-review witness prep.

Source: EDRM Search Glossary.

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Extended Industry Standard Architecture

See: EISA (Extended Industry Standard Architecture)

Extensible Markup Language (XML)

Code which describes the content of data.

A subset of SGML that is used to describe the structure and content of documents. The "extensible" part of its name indicates that it can be used to create new data structures, which makes it more powerful than HTML.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

HTML
Java
JavaScript
SGML
SGML/HyTime

Extensions/Sizes Filter

A filter option that allows for including/excluding files with embedded objects, for including or excluding certain file types or non-mail e-mail items (such as Calendar, Appointments, Tasks, Contacts, etc), as well as establishing thresholds for file size.

Source: Ibis Consulting, Glossary.

See also:

Date filter
Filter
MD5-known filter
Sender/recipient filter

External Drive

See: Portable drive

Extranet

An intranet connection that is made accessible to authorized users outside of the network.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

An intranet to which the owners provide limited access to outside users such as clients or co-counsel.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

An Internet based access method to a corporate intranet site by limited or total access through a security firewall. This type of access is typically utilized in cases of joint venture and vendor client relationships.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Extrinsic Data

Information about a file, such as file signature, author, size, name, path, and creating and modification dates. This data is the accumulation of what is in the file, on the media label, discovered by the operator, and contributed by the user. Collectively, it represents one of the values of examining an electronic file as opposed to the printed version.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

See also:

Customer-added metadata
Document metadata
Email metadata
File parameters
File system metadata
File-specific metadata
General metadata
Metadata
Vendor-added metadata

F

F

van Rijsbergen's F. A formula for combining precision and recall into a single number to make it easier to compare the information retrieval accuracy of different systems.

Source: Herb Roitblat, Predictive Coding Glossary.

F1

The Harmonic Mean of Recall and Precision, often used in Information Retrieval studies as a measure of the effectiveness of a search or review effort, which accounts for the tradeoff between Recall and Precision. In order to achieve a high F1 score, a search or review effort must achieve both high Recall and high Precision.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

One form of van Rijsbergen's F formula for combining precision and recall into a single number to make it easier to compare the information retrieval accuracy of different systems. F1 is the weighted harmonic mean of precision and recall = *2 * precision * recall / (precision + recall)*.

Source: Herb Roitblat, Predictive Coding Glossary.

Faceted Query

A search query wherein the system returns a set of alternative values that are typically subcategories of the original query. For example, on a website selling television sets, a user might enter an initial query consisting of the word "TV." The system will then present a list of various subclasses of TVs, categorized, for example, by the size of the screen. In eDiscovery, a system might return a list of email senders in response to an initial query. A faceted query is a type of query expansion, where the expanded queries are categorized and can be selected.

Source: Herb Roitblat, Search 2020: The Glossary.

Facsimile

A process of transmitting documents by scanning them to digital, converting to analog, transmitting over phone lines and reversing the process at the other end and printing. "Group 3" indicates the 3rd generation of faxes which transmits a page at 9600 baud in about a minute – with a normal resolution of 203 x 98 dpi and a fine resolution of 203 x 196.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

False Negative (FN)

A Relevant Document that is Missed (i.e., incorrectly identified as Non-Relevant) by a search or review effort. Also known as a Miss.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

One of four response states in a categorization task. True positive responses are those that are truly in the positive category and are classified as negative.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

False Positive (FP) True Negative (TN) True Positive (TP)

False Negative Rate (FNR)

The fraction (or Proportion) of Relevant Documents that are Missed (i.e., incorrectly identified as Non-Relevant) by a search or review effort. Note that False Negative Rate + Recall = 100%, and that 100% – Recall = False Negative Rate.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

False Positive (FP)

A Non-Relevant Document that is incorrectly identified as Relevant by a search or review effort.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

One of four response states in a categorization task. False positive responses are those that are truly in the negative category and are classified as positive.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

False Negative (FN) True Negative (TN) True Positive (TP)

False Positive Rate (FPR)

The fraction (or Proportion) of Non-Relevant Documents that are incorrectly identified as Relevant by a search or review effort. Note that False Positive Rate + True Negative Rate = 100%, and that 100% – True Negative Rate = False Positive Rate. In Information Retrieval, also known as Fallout.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Family Range

A family range describes the range of documents from the first Bates production number assigned to the first page of the top most parent document through the last Bates production number assigned to the last page of the last child document.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Family Relationship

A family relationship is formed among two or more documents that have a connection or relatedness because of some factor.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Fast Mode Parallel Port

See: Enhanced Parallel Port (EPP)

FAT (File Allocation Table)

An internal data table on DOS-based disks that lists the contents and address of each file on the disk.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

File system NT filing system NTFS

Fax Board

An adapter that is installed inside a computer to allow direct faxing.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Feature

A characteristic of an item. In text, a feature is usually a word, but it could be a phrase or other grouping of words. In a search engine, features are the items that are specifically indexed.

Source: Herb Roitblat, Search 2020: The Glossary.

Feature Engineering

The process of identifying Features of a Document that are used as input to a Machine Learning Algorithm. Typical Features include words and phrases, as well as metadata such as subjects,

dates, and file types. One of the simplest and most common Feature Engineering techniques is Bag of Words. More complex Feature Engineering techniques include the use of Ontologies and Latent Semantic Indexing.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Features

The units of information used by a Machine Learning Algorithm to Classify or Prioritize Documents. Typical Features include text fragments such as words or phrases, and metadata such as sender, recipient, and sent date. See also Feature Engineering.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Fiber Optics

Transmitting with light pulses over cables made from thin strands of glass.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Field

A name for an individual piece of standardized data to be extracted from an image collection. Fields can be the author of a document, a recipient, the date of a document or any other piece of data common to most documents in an image collection.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A unit of information in a database. Database records, for example, consist of an ordered list of fields where a specific kind of information is stored in each field. Fields are often printed as columns in database reports.

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient

Subject category Summary Text

Field Separator

A code, usually a comma, that separates the fields in a record. (Also, a delimiter.)

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Fielded Search

Fielded searches are based on values stored as metadata rather than actual content of an electronic asset. Searches can be refined using metadata information extracted during processing, such as sender or receiver, creation date, modified date, author, file type and title, as well as subjective user-defined values that may be ascribed to a document as part of downstream review. See also Parametric Search.

Source: EDRM Search Glossary.

File

A document or program as well as a unit of storage or file management. Each file is a set of bytes (each byte typically consists of 8 bits) that is stored on some media, or inside an archive. Files can be transmitted over communication lines using communication protocols such as SMTP/POP3 (Mail), FTP, HTTP. Files may (or may not) have different attributes (metadata). There are many different types of files: data files, text files, program files, directory files, and so on. Different types of files store different types of information. For example, program files store programs, whereas text files store text.

Source: Ibis Consulting, Glossary.

An element of data storage in a file system. A collection of data or information that has a name, called the filename. Almost all information stored in a computer must be in a file. There are many different types of files: data files, text files, program files, directory files, and so on.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

In word processing, a piece of text that is usually one document long.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

In a database, a complete collection of records treated as one unit.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A collection of logically related data records.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A collection of data of information stored under a specified name on a disk.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

File Allocation Table

See: FAT (File Allocation Table)

File Compression

A technology for storing data in fewer bits, it makes data smaller so less disk space is needed to represent the same information. Compression programs like WinZip and UNIX compress are valuable to network users because they save both time and bandwidth. Data compression is also widely used in backup utilities, spreadsheet applications, and database management systems.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

A technology for storing data in fewer bits, it makes data smaller so less disk space is needed to represent the same information. Data compression is widely used to backup utilities, spreadsheet applications, and database management systems. Compressed files must be decompressed in order to be useable.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A technology that reduces the size of a file.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Any method which reduces the amount of data necessary to transmit information from one point to another. Compression generally eliminates redundant information and/or predicts where changes will occur. "Lossless" compression techniques totally preserve the integrity of the input. "Lossy" methods disregard some of the originals. The ratio of the file sizes of a compressed file to an uncompressed file, e.g., with a 20:1 compression ratio, an uncompressed file of 1MB is compressed to 50KB.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A technology that reduces the size of a file. Compression programs are valuable to network users because they help save both time and bandwidth.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

File Extension

Distinguishes a file's format for the application used to create the file and can be used to simplify the process of locating data.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A tag of three or four letters, preceded by a period, which identifies a data file's format or the application used to create the file. File extensions can streamline the process of locating data. For example, if one is looking for incriminating pictures stored on a computer, one might begin with the .gif and .jpg files.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The last (typically 3 characters) following a period in a file name that indicates what kind of file it is. MS word documents typically end with ".doc," etc. File extensions are used by the operating system to determine the default application to use to open a file.

In DOS and some other operating systems, one or several letters at the end of a filename. Filename extensions usually follow a period (dot) and indicate the type of information stored in the file. For example, in the filename LETTER.DOC, the extension is DOC, which indicates that the file is a word processing file.

File Parameters

File data which can be read from the file system, including folder location name, file name, creation date, last modified date, last accessed date, and file size.

Source: Ibis Consulting, Glossary.

See also:

Customer-added metadata	Extrinsic data	General metadata
Document metadata	File system metadata	Metadata
Email metadata	File-specific metadata	Vendor-added metadata

File Server

Utilized when many computer systems are connected together as part of a LAN, a file server can retain email messages, financial data, word processing information, or be used to backup the network.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A computer that is the central storage unit for a local area network (LAN).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

When several or many computers are networked together in a LAN situation, one computer may be utilized as a storage location for files for the group. File servers may be employed to store e-mail, financial data, word processing information or to back-up the network.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Computer	Minicomputer	Workstation
Laptop computer	Notebook computer	
Microcomputer	Personal computer	

File Set

Any group of files to be processed, either in or outside of a container.

Source: Ibis Consulting, Glossary.

File Sharing

Sharing of computer data or space on a network. File sharing allows multiple users to use the same file by being able to read, modify, copy and/or print it. File sharing users may have the same or different levels of access privilege.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

One of the key benefits of a network is the ability to share files stored on the server among several users.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

File Signature

Within a file, the file signature is the information about the true program-related origin of the file, and therefore, its type. Tools for reading file signatures identify the true program source, even if the file extension has been changed.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

File System

The system that an operating system or program uses to organize and keep track of files. For example, a hierarchical file system is one that uses [Directory|directories}] to organize files into a tree structure. Types of file systems include file allocation table (FAT) and Windows NT file system (NTFS).

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

See also:

FAT	NT filing system	NTFS

File System Metadata

Data that can be obtained or extracted about a file from the file system storing the file. Contrast with document metadata and email metadata.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

See also:

Customer-added metadata	Extrinsic data	General metadata
Document metadata	File parameters	Metadata
Email metadata	File-specific metadata	Vendor-added metadata

File Transfer Protocol (FTP)

The protocol for exchanging files over the Internet. FTP works in the same way as HTTP for transferring Web pages from a server to a user's browser and SMTP for transferring electronic mail across the Internet -- in that, like these technologies, FTP uses the Internet's TCP/IP protocols to enable data transfer.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Webopedia Computer Dictionary, http://www.pcwebopedia.com/TERM/F/FTP.html.

An Internet protocol that enables you to transfer files between computers on the Internet.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Applied Discovery's Glossary, http://www.nysd.uscourts.gov/courtweb/pdf/D02NYSC/03-04265.PDF#page=24

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A protocol used on the internet for exchanging files.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

An Internet protocol to move files from one computer to another.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

File-Specific Metadata

File-specific metadata is defined as data about the file itself, such as title, subject, author, keywords, comments, etc. for application files and MessageID, header, text, time received, number of attachments, etc. for mail items.

Source: Ibis Consulting, Glossary.

See also:

Customer-added metadata	Extrinsic data	General metadata
Document metadata	File parameters	Metadata
Email metadata	File system metadata	Vendor-added metadata

Filename

The name given to a computer file in order to distinguish it from other files; may contain an extension that indicates the type of file.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The name of a file. All files have names. Different operating systems impose different restrictions on filenames. Most operating systems, for example, prohibit the use of certain characters in a filename and impose a limit on the length of a filename. In addition, many systems, including DOS and UNIX, allow a file name extension that consists of one or more characters following the proper file name. The filename extension usually indicates what type of file it is.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

The name of a file. All files have names. Different operating systems impose different restrictions on filenames. Most operating systems, for example, prohibit the use of certain characters in a filename and impose a limit on the length of a filename. In addition, many systems, including DOS and UNIX, allow a file name extension that consists of one or more characters following the proper file name. The filename extension usually indicates what type of file it is. The line of usually Roman alphabet characters plus a file attribute's FILE EXTENSION (zip, doc, info, xwy) separated by the symbol '.' or a file attribute's FILE EXTENSION (zip, doc, info, xwy). This may (or MAY NOT) be an indicator of FILE FORMAT FILE FORMAT. File format specification allows different applications to understand the same files.

Source: Ibis Consulting, Glossary.

Filename Extension

In DOS and some other operating systems, one or several letters at the end of a filename. Filename extensions usually follow a period (dot) and indicate the type of information stored in the file. For example, in the filename LETTER.DOC, the extension is .DOC, which indicates that the file is a word processing file.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

Filter

Various methods of reducing a data set.

Source: Ibis Consulting, Glossary.

See also:

Date filter

Extensions/sizes filter

MD5-known filter

Sender/recipient filter

Filtering

The method of reducing a data set by applying one or more filters.

Source: Ibis Consulting, Glossary.

Electronic filtering of emails and files for privilege or by keyword, file, type or name. Filtering removes files that do not fit the search criteria and reduces the volume of data that requires further investigation.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Find Similar

A search method that identifies Documents that are similar to a particular exemplar. Find Similar is commonly misconstrued to be the mechanism behind Technology-Assisted Review.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Fingerprinting

See: Hash

Firewall

A system intended to thwart unauthorized access to or from a private network that is often used to prevent unauthorized users from accessing private networks connected to the internet.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A set of related programs that protect the resources of a private network from users from other networks.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Flash Drive

See: Jump Drive

Flat File Database

A database with all data in a single list, similar to a telephone book or a Rolodex.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Database
Full text database
Relational database
SQL
WAIS - wide area information server

Flatbed Scanner

A scanner design in which the document is placed in the scanner's bed, either manually or by an automatic document feeder, and remains stationary during scanning. As a result, flatbed scanners provide a more stable target than other scanner designs, but they are generally slower.

Source: RSI, Glossary.

A flat-surface scanner that allows users to input books and other documents.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Double-sided scanner
Duplex scanner
Scanner
Simplex scanner

Floppy Disk

A thin magnetic film disk that is used as an older method for storing data.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Small removable disks, also known as diskettes, that come in two sizes, 3.5” and 5.25”. The amount of data that can be stored on a diskette depends on the size, and can be 360 kilobytes to 1.4 megabytes.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

An increasingly rare storage medium consisting of a thin magnetic film disk housed in a protective sleeve.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

CD
CD-R
CD-ROM
CD-RW
Disc
Disk

Diskette	Jaz disk	Optical disk
DVD	Laser disc	Storage media
DVD-ROM	Magnetic disk	WORM disk
Hard disk	Magnetic storage media	Zip disk
Hard drive	Media	

Floppy Disk Drive

A Floppy Disk Drive, also called FDD or FD for short, is a computer disk drive that enables a user to save data to removable diskettes. Although 8" disk drives were first made available in 1971, the first real disk drives used were the 5 1/4" floppy disk drives, which were later replaced with the 3 1/2" floppy disk drives.

Source: Computer Hope, FDD definition, http://www.computerhope.com/jargon/f/fdd.htm

See also:

Disk drive	Portable drive	Zip drive
Jaz drive	Storage device	
Magneto-optical drive	Tape drive	

FN

See: False Negative (FN)

Fog Computing

Using a combination of local and cloud computing resources. Data may be kept locally for security reasons, to avoid the expense and time of moving them to centralized cloud storage, or to meet compliance requirements.

Source: Herb Roitblat, Search 2020: The Glossary.

Folder Browser

A system of on-screen folders (usually hierarchical or “stacked”) used to organize documents. For example, the File Manager program in Microsoft Windows is a type of folder browser that displays the directories on your disk.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Font

A complete set of characters in a distinctive type style and size.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Forensic Analysis

The scientific analysis of computer media for legal reasons. Typically forensic analysis is intended to discover whether responsive documents or other data have been deleted from a machine. Special software, equipment, and techniques are used to detect hidden information. Such investigations have been more common in criminal proceedings than in litigation. Strict protocols must be employed to avoid evidence spoliation.

See also:

Computer evidence
Computer forensics
Computer investigations
Discovery
Electronic discovery / e-discovery
Electronic evidence
Forensics
Mirroring

Forensic Copy

An exact bit-by-bit copy of the entire physical hard drive of a computer system, including slack and unallocated space.

Source: Merrill Corporation, Electronic Discovery Glossary.

A precise bit-by-bit copy of a computer system's hard drive, including slack and unallocated space.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Bitstream copy
Image
Imaged copy
Mirror image

Forensically Sound Procedures

Procedures used for acquiring electronic information in a manner that ensures it is “as originally discovered” and is reliable enough to be admitted into evidence. Such procedures are defined in part by the US Department of Justice publication “Searching and Seizing Computers and Obtaining Electronic Evidence in Criminal Investigations,” http://www.usdoj.gov/criminal/cybercrime/s&smanual2002.htm.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

See also:

Chain of custody
Chain of evidence

Forensics

For electronic data, the discovery discipline that includes the physical acquisition of digital data using methodology that satisfies evidentiary requirements of chain-of-custody and authentication. Preserving the evidence includes performing code and encryption cracking, searching and retrieving elusive data, determining if files have or have not been deleted, recovering deleted files, and determining use, including Internet, network access, printing, filing, and copying.

Source: Ibis Consulting, Glossary.

In document management terms, forensic work is comprised of:

- Recreating "deleted" or missing files from hard drives
- Validating dates and logged in authors / editors of documents
- Certifying key elements of documents and/or hardware for legal purposes

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Computer evidence	Discovery	Electronic evidence
Computer forensics	Electronic discovery / e-discovery	Forensic analysis
Computer investigations		Mirroring

Formal Search

Formal search includes executing, tracking. reporting and measure impact, and iterate through sets of multiple logical queries. See also, Iterative Search.

Source: EDRM Search Glossary.

Format

The internal structure of a file, which defines the way it is stored and used. Specific applications may define unique formats for their data (i.e., "MS Word document file format"). Many files may only be viewed or printed using their originating application or an application designed to work with compatible formats. Computer storage systems commonly identify files by a naming convention that denotes the format (and therefore the probable originating application) (i.e., "DOC" for Microsoft Word document files; "XLS" for Microsoft Excel spreadsheet files; "TXT" for text files; and "HTM" (for Hypertext Markup Language (HTML) files such as Web pages). Users may choose alternate naming conventions, but this may affect how the files are treated by applications.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The organization of data on a disk.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

To prepare a disk for use. Formatting a disk consists of erasing old information on the disk and adding new codes to control information recording.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Forms Processing

A specialized imaging application designed for handling pre-printed forms. Forms processing systems often use high-end (or multiple) OCR engines and elaborate data validation routines to extract hand-written or poor quality print from forms that go into a database. This type of imaging application faces major challenges, since many of the documents scanned were never designed for imaging or OCR.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Forms Routing

The process of routing a form throughout an organization electronically – without any paper copies.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Formula Report

A report listing the spreadsheet formulas cell by cell. These formulas describe how a computation was performed in calculating the values displayed in a spreadsheet.

FP

See: False Positive (FP)

FPR

See: False Positive Rate (FPR)

Fragmented Data

Live data that has been disseminated and stored in multiple areas on a single hard drive or disk.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Fragmented data is live data that has been broken up and stored in various locations on a single hard drive or disk.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

"Live" data that has been broken up and stored in various locations on a single hard drive. Most files are stored this way.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ambient data	Residual data	Swap file
Free space	Slack space	Unallocated space

Free Space

Unused clusters on a hard disk.

Source: PCMag, definition of free space, http://www.pcmag.com/encyclopedia/term/56700/free-space

See also:

Ambient data	Residual data	Swap file
Fragmented data	Slack space	Unallocated space

Frequency Analysis

Utilized iteratively throughout the life cycle of a project as search criteria are modified, frequency analysis may be used to evaluate the effectiveness of the initial search criteria. The search terms are tested to determine whether they effectively discriminate between potentially relevant and clearly non-relevant data. Frequency analysis is a reality check on the search results versus the overall collection size and the reasonably expected proportion of relevant results. It does not address the recall or completeness of relevant items out of the collection.

Source: EDRM Search Glossary.

FRN

See: False Negative Rate (FNR)

Front-End / Back-End

Expressions that describe programs relative to the user. A front-end program is one that users interact with directly, while a back-end program supports the front-end services.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

FTP

See: File Transfer Protocol (FTP)

Full Duplex

Data communications devices which allow full speed transmission in both directions at the same time.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Full Text Database

A database in which the entire text of documents is electronically available for searching by keywords or phrases using Boolean logic.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Database
Flat file database
Relational database
SQL
WAIS - wide area information server

Full Text Indexing

Enables the retrieval of documents by either their word or phrase content. Every word in the document is indexed into a master word list with pointers to the documents and pages where each occurrence of the word appears.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Full Text Search

A filter option that allows for including/excluding files that have been searched for designated terms or phrases (either uploading a list in any text file format or created on-the-spot). General metadata, file names, and body text are searched.

Source: Ibis Consulting, Glossary.

The ability to search a data file for specified key(s) defined by the occurrence of words, numbers and/or combinations or patterns thereof.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

The ability to search all of the words of a document, not just those contained in special fields, metadata, codes, or summaries.

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search

Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Function Key

A key on the keyboard that controls specialized functions other than normal typing. Function keys include F1 through F10, CTRL, ALT, SHIFT, PAGE UP, PAGE DOWN, DELETE, and INSERT.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Fuzzy Search

Fuzzy search allows searching for word variations such as in the case of misspellings. Typically, such searching includes some form of distance and score computations between the specified word and the words in the corpus.

Source: EDRM Search Glossary.

A search technique that identifies ESI based on terms close to another term, with closeness defined as a typographical difference and/or change. For example, snitch, switch, and swanky can all match swatch, depending on how many incorrect letters are allowed within the search threshold.

Source: EDRM Search Guide Glossary.

Search that locates words closely match the spelling of the primary word.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A full-text search procedure that looks for exact matches as well as similarities to the search criteria, in order to compensate for spelling or OCR errors.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search

Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

G

Gain Curve

A graph that shows the Recall that would be achieved for a particular Cutoff. The Gain Curve directly relates the Recall that can be achieved to the effort that must be expended to achieve it, as measured by the number of Documents that must be reviewed and Coded.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Garbage In, Garbage Out (GIGO)

Well-known computer adage which refers to the fact that the contents of a database are only as good as the data originally entered. Data entered incorrectly will not provide accurate search results and will lead users to rely on incorrect information.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Gaussian Distribution

A normal distribution. A bell shaped curve representing the likelihood of different values as they depart from the average.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

Normal Distribution

Gaussian Estimate

A Statistical Estimate of a Population characteristic using Gaussian Estimation. It is generally expressed as a Point Estimate accompanied by a Margin of Error and a Confidence Level, or as a Confidence Interval accompanied by a Confidence Level.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

GB (Gigabyte)

Approximately one billion bytes. Often shortened to "gigs" or GB.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

One billion bytes. Also expressed as one thousand megabytes. In terms of image storage capacity, one gigabyte equals approximately 17,000 81/2" x 11" pages scanned at 300 dpi, stored as TIFF Group IV images.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

The equivalent of one billion (actually 1,073,741,824) bytes; or one million kilobytes, or one thousand megabytes.

Equal to 1,000 megabytes (MB) or 1,073,741,824 bytes.

See also:

Bit	MB - megabyte	EB - exabyte
Byte	TB - terabyte	
KB - kilobyte	PB - petabyte	

General Metadata

General metadata is not real metadata, but is data about a file other than the contents. This includes data such as original name, date created, date last modified, file extension, file type, file size, MD5, etc. for application files and folder name, size, subject, time created, MD5, etc. for mail items. File-specific metadata is defined as data about the file itself, such as title, subject, author, keywords, comments, etc. for application files and MessageID, header, text, time received, number of attachments, etc. for mail items.

Source: Ibis Consulting, Glossary.

See also:

Customer-added metadata	Extrinsic data	File-specific metadata
Document metadata	File parameters	Metadata
Email metadata	File system metadata	Vendor-added metadata

GHz (Gigahertz)

When referring to a computer processor or CPU, GHz is a clock frequency, also known as a clock rate or clock speed, representing a cycle of time. An oscillator circuit supplies a small amount of electricity to a crystal each second that is measured in KHz, MHz, or GHz. "Hz" is an

abbreviation of Hertz, and "K" represents Kilo (thousand), "M" represents Mega (million), and "G" represents Giga (thousand million).

Source: Computer Hope, GHz definition, http://www.computerhope.com/jargon/g/ghz.htm

See also:

Hz KHz MHz

GIF (Graphic Interchange File)

A bit-mapped graphics file format used by the on the Internet. GIF supports color and various resolutions. It also includes data compression, but because it is limited to 256 colors, it is more effective for scanned images such as illustrations rather than color photos.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

An image storage format that is widely used on the web.

Source: RSI, Glossary.

A compressed file format used by the CompuServe system for photographs. Limited to 256 colors.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A computer compression format for pictures.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Image file format
Joint photographic expert group
JPEG
Multi-page TIFF
PDF
PNG
Portable Document Format
Portable network graphic
Searchable TIFF
Single-page TIFF
TIFF

Gigabyte

See: GB (Gigabyte)

Gigahertz

See: GHz (Gigahertz)

GIGO

See: Garbage In, Garbage Out (GIGO)

Global Aerospace

Global Aerospace Inc. v. Landow Aviation, Consol. Case No. CL 61040, 2012 WL 1431215 (Va. Cir. Ct. Apr. 23, 2012). The first State Court Order approving the use of Predictive Coding by the producing party, over the objection of the requesting party, without prejudice to the requesting party raising an issue with the Court as to the completeness or the contents of the production, or the ongoing use of Predictive Coding. The order was issued by Loudoun County Circuit Court Judge James H. Chamblin.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Global Deduplication

Deduplication of Documents across multiple custodians. Also referred to as Horizontal Deduplication. (*Cf*. Vertical Deduplication.)

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

See also:

Basic de-duplication	De-duplication	Horizontal Deduplication
Case de-duplication	Duplicate	Production de-duplication
Custodian de-duplication	Dynamic de-duplication	Vertical Deduplication

Gold Standard

The best available determination of the Relevance or Non-Relevance of all (or a sample) of a Document Population, used as benchmark to evaluate the effectiveness of a search and review effort. Also referred to as Ground Truth.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Goodhart's Law

An observation made in 1975 by Charles Goodhart, Chief Adviser to the Bank of England, that statistical economic indicators, when used for regulation, become unreliable. Restated and generalized in 1997 by University of Cambridge Professor Marilyn Strathern as "When a measure becomes a target, it ceases to be a good measure." Within the context of Electronic Discovery, Goodhart's Law suggests that the value of Information Retrieval measures such as Recall and Precision may be compromised if they are prescribed as the definition of the reasonableness of a search or review effort.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Graphic Interchange File

See: GIF (Graphic Interchange File)

Graphical User Interface (GUI)

Abbreviated GUI (pronounced GOO-ee). A program interface that takes advantage of the computer's graphics capabilities to make the program easier to use. Well-designed graphical user interfaces can free the user from learning complex command languages.

Source: http://www.webopedia.com/TERM/G/Graphical_User_Interface_GUI.html

Elements include such things as windows, icons, buttons, cursors, and scroll bars.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Software programs that use special icons and other symbols to assist in performing functions, decrease reliance on keyboard skills, and reduce training time. The two most prominent examples are the Apple interface and Microsoft Windows. Pronounced "gooey."

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Graphical User Interface, or "gooey". Presenting an interface to the computer user comprised of pictures and icons, rather than words and numbers.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A set of screen presentations and metaphors that utilize graphic elements such as icons in an attempt to make an operating system easier to use.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Graphics Board

A board that allows the screen to display graphic images.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Grayscale

An image type that uses black, white, and a ranges of shades of gray. The number of shades of gray depends on the number of bits per pixel. The larger the number of shades of gray, the better the image will look, and the larger the file will be.

Source: RSI, Glossary.

The use of many shades of gray to represent an image. Continuous-tone images, such as black-and-white photographs, use an almost unlimited number of shades of gray. Conventional computer hardware and software, however, can only represent a limited number of shades of gray (typically 16 or 256). Gray-scaling is the process of converting a continuous-tone image to an image that a computer can manipulate. The binary range of a graphic representation between pure black and pure white. A scale of 256 shades of gray will be a better representation than 16 shades.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Groupware

Software designed to promote action among members of specific groups within an organization. The best-known groupware is Lotus Notes.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Software designed to operate on a network and allow several people to work together on the same documents and files.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

GUI

See: Graphical User Interface (GUI)

H

Hadoop

A software library that provides a framework for distributed processing of large data sets. Hadoop allows complex processes to be broken down into basic computing tasks, which can be distributed among a potentially large number of individual computers. The results of these distributed computations can then be merged to obtain the final product.

Source: Herb Roitblat, Search 2020: The Glossary.

Half Duplex

Transmission systems which can send and receive, but not at the same time.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Halftone

The graphic representation of an object by dots, which simulate continuous tones. Usually used to represent or replicate an original photograph input.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Halftone Dot

Vary in size; larger appear darker, smaller appear lighter.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Hard Disk

Internal hardware which stores and provides access to large amounts of information. Most new computers include an internal hard disk that contains several gigabytes of storage capacity.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A high-capacity magnetic media storage device, also known as the "fixed disk." Hard disks are either internal or external. An internal hard disk can be used only with the computer in which it is installed, while an external hard disk can be moved from one computer to another.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

CD	DVD	Magnetic storage media
CD-R	DVD-ROM	Media
CD-ROM	Floppy disk	Optical disk
CD-RW	Hard drive	Storage media
Disc	Jaz disk	WORM disk
Disk	Laser disc	Zip disk
Diskette	Magnetic disk	

Hard Drive

The primary computer storage medium in desktop and laptop computers.

Source: RenewData, Glossary (10/5/2005).

The primary hardware that a computer uses to store information, typically magnetized media on rotating discs.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

The primary storage unit on PCs, consisting of one or more magnetic media platters on which digital data can be written and erased magnetically.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A magnetic storage device usually inside a computer that stores files. Hard drive capacities are usually measured in gigabytes (GB). In addition to hard drives, computers often contain CD drives and floppy disk drives. When you save a file, it is usually stored on the computer's hard drive.

A high-capacity magnetic media storage device, also known as the "fixed disk." Hard disks are either internal or external. An internal hard disk can be used only with the computer in which it is installed, while an external hard disk can be moved from one computer to another.

Internal hardware which stores and provides access to large amounts of information. Most new computers include an internal hard disk that contains several gigabytes of storage capacity.

See also:

CD	DVD	Magnetic storage media
CD-R	DVD-ROM	Media
CD-ROM	Floppy disk	Optical disk
CD-RW	Hard disk	Storage media
Disc	Jaz disk	WORM disk
Disk	Laser disc	Zip disk
Diskette	Magnetic disk	

Hardcopy

The paper version of a document.

Hardware

All the mechanical and electrical parts of a computer.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Hardware Key

External security used with some software. Without this key, the software will not function.

Source: RSI, Glossary.

Harmonic Mean

The reciprocal of the average of the reciprocals of two or more quantities. If the quantities are named a and b, their Harmonic Mean is $\frac{2}{\frac{1}{a}+\frac{1}{b}}$. In Information Retrieval, F1 is the Harmonic Mean of Recall and Precision. The Harmonic Mean, unlike the more common arithmetic mean (i.e., average), falls closer to the lower of the two quantities. As a summary measure, a Harmonic Mean may be preferable to an arithmetic mean because a high Harmonic Mean depends on both high Recall and high Precision, whereas a high arithmetic mean can be achieved with high Recall at the expense of low Precision, or high Precision at the expense of low Recall.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Hash

An algorithm that creates a value to verify duplicate electronic documents. A hash mark serves as a digital thumbprint.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

See also:

Hash value

Hashing / Hash / Hash Value

MD5

SHA-1

Hash Value

A unique, identifying number of a file calculated by a hash algorithm, e.g. MD5.

Source: Ibis Consulting, Glossary.

A computed numerical value that represents a “digest” of the content of a file. If and only if two documents are identical to the letter will they return the same hash value. The Hash value is used as part of a digital signature and to compare document content in the de-duping process.

See also:

Hash

Hashing / Hash / Hash Value

MD5

SHA-1

Hashing / Hash / Hash Value

A statistical method used to reduce the contents of a Document to a single, fixed-size, alphanumeric value, which is, for all intents and purposes, unique to a particular Document; the single, fixed-size alphanumeric value resulting from Hashing a particular Document. Common

Hashing Algorithms include, but are not limited to, MD5, SHA-1, and SHA-2. Hashing and Hash Values are typically used for Document identification, Deduplication, or ensuring that Documents have not been altered.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

See also:

Hash MD5

Hash value SHA-1

HD (High Density)

High density floppy disks; a 5.25" holds 1.2 MB and a 3.5" holds 1.4 MB.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Help

Instructions that assist a user on how to set up and use a product including but not limited to software, manuals and instruction files.

Source: RSI, Glossary.

Hertz (Hz)

Cycles per second. Often used with metric prefixes, as in kiloHertz (kHz).

Source: RSI, Glossary.

See also:

KHz MHz GHz

Heuristic

A general practical approach to solving a problem that is useful to address the problem, but whose result is not guaranteed. Examples of heuristics include mental shortcuts, rules of thumb, and general strategies. Unlike an algorithm, a heuristic is not guaranteed to produce a specific result.

Source: Herb Roitblat, Search 2020: The Glossary.

Hexadecimal

A number system with a base of 16 (2 4), 4 bits. The position digits are 0-9, A-F, where F equals the decimal value, 15.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Hierarchical Storage Management (HSM)

Software that automatically migrates files from on-line to near-line storage media, usually on the basis of the age or frequency of use of the files.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

High Density

See: HD (High Density)

Hit

A term to describe the results of a search query. A search for a specific name may produce twenty "hits," which means the name appears twenty times in the database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Hit / Hit Level Results

To arrange or designate according to categorization such as potentially responsive or privileged versus non-responsive or not-privileged.

Source: EDRM Search Glossary.

Hold

See: Legal Hold

Holorith

Encoded data on aperture cards or old-style punch cards that contained encoded data.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Horizontal Deduplication

See Global Deduplication. (*Cf*. Vertical Deduplication.)

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

See also:

Basic de-duplication
Case de-duplication
Custodian de-duplication
De-duplication
Duplicate
Dynamic de-duplication
Global Deduplication
Production de-duplication
Vertical Deduplication

Host

Computer on which an application or database resides.

Source: RSI, Glossary.

In a network, the central computer which controls the remote computers and holds the central databases.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Hosting

Providing an application on-line (see Application Service Provider) for one or more clients, usually housed in a data center.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Hot Key

An individual key that is programmed to perform a specific function.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

HP-PCL

A Hewlett-Packard graphics file format.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

HPGL

A Hewlett-Packard graphics file format.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

HSM

See: Hierarchical Storage Management (HSM)

HTML (Hypertext Markup Language)

A set of codes inserted into a file or document that is intended to display through a Web browser. HTML tells the browser how to display a document's words and images as a Web page. Each markup symbol is referred to as an element or a tag.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A language that uses tags to structure text into headings, paragraphs, lists and links. It tells a Web browser how to display text and images.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

The underlying program structure of text on the World Wide Web.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Developed by CERN of Geneva, Switzerland. The document standard of choice of Internet. (HTML+ adds support for multi-media.) Use in internet application.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

The tag-based ASCII language used to create pages on the Web.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The standard code used to tell a browser how a web page should be displayed.

See also:

Java
JavaScript
SGML
SGML/HyTime
XML

Hub

A central unit that repeats and/or amplifies data signals being sent across a network.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Hypertext Markup Language

See: HTML (Hypertext Markup Language)

Hz

See: Hertz (Hz)

I

IBM Watson Project

A pioneering project from IBM that implements what they call cognitive computing. Watson was originally built to demonstrate its advanced natural language processing, knowledge representation, machine learning, and information retrieval processes in the context of the

game Jeopardy. Its speed an extensive knowledge base, along with its automated reasoning allowed it to win the competition against human Jeopardy players. The same computational framework has since been applied on a commercial basis to other open-ended question answering tasks, including medical diagnosis, and recipe construction.

Source: Herb Roitblat, Search 2020: The Glossary.

I/O (Input/Output)

The transfer of information in and out of a computer's memory.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Icon

A graphic image or picture of a program or task designed to represent that program or task.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

In a GUI, a picture or drawing which is activated by "clicking" a mouse to command the computer program to perform a predefined series of events.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

ICR (Intelligent Character Recognition)

The conversion of scanned images (bar codes or patterns of bits) to computer recognizable codes (ASCII characters and files) by means of software/programs which define the rules of and algorithms for conversion.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Dirty OCR

OCR

Optical Character Recognition

Pattern recognition

IDE (Integrated Drive Electronics)

An engineering standard for interfacing PC's and hard disks.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

IEEE (Institute of Electrical and Electronic Engineers)

An international association which sponsors meetings, publishes a number of journals and establishes standards.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

IM (Instant Messaging)

Instant Messaging is a form of electronic communication which involves immediate correspondence between two or more users who are all online simultaneously.

Source: Merrill Corporation, Electronic Discovery Glossary.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Electronic communication allowing for instant correspondence between two or more users who are online at the same time.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Image

A bit-by-bit duplicate of a backup tape or hard drive that is forensically sound. Also known as an image copy, a forensic copy or a mirror image.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

As distinct from document imaging, electronic evidence is making an identical copy of a hard drive. Also known as a “mirror image” or “mirroring”.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

An electronic “picture” of how the document would look if printed. Images can be stored in various file formats, the most common of which are TIFF and PDF.

See also:

Bitstream copy
Forensic copy
Imaged copy
Mirror image

Image Compression Board

An imaging-dedicated processor. Relieves the CPU (Central Processor Unit - the computer's main chip) from many imaging-specific tasks - compression, decompression, display, zooming, shrinking, scale-to-gray. In fact, does them better than the CPU.

Source: RSI, Glossary.

Image Enable

A software function that creates links between existing applications and stored images.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Image File Format

When a page is scanned, the page can be stored in a number of file types. The type should be chosen based on the desired use of the image, and the software that will be used. Different file formats commonly use different methods of compression as well, and some types of images compress better using some formats rather than others.

Source: RSI, Glossary.

See also:

GIF
Graphic Interchange File
Joint photographic expert group
JPEG
Multi-page TIFF
PDF
PNG
Portable Document Format
Portable network graphic
Searchable TIFF
Single-page TIFF
TIFF

Image Key

The name of a file created when a page is scanned in a collection.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Image Processing

To capture an image or representation, enter in a computer and process and manipulate it.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Think of "data processing": it refers to the manipulation of raw data to solve some problem or enlighten the user in some way not possible without manipulation.

Source: RSI, Glossary.

Image Processing Card (IPC)

A board mounted in either the computer, scanner or printer that facilitates the acquisition and display of images. The primary function of most IPCs is the rapid compression and decompression of image files.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Image Resolution

The fineness or coarseness of an image as it was digitized, measured as dots or pixels per inch. All other things being equal, the higher the resolution, the better is the image. Resolutions of 200 dots per inch (dpi). The higher the resolution, the greater the amount of detail that can be shown.

Imaged Copy

A "mirror image" bit-by-bit copy of a hard drive, i.e. a complete replication of the physical drive regardless of how the drive is organized or whether the image created contains meaningful data in whole or in part. From an imaged copy of a hard drive it is possible to reconstruct the entire contents and organization of the source drive from which it was taken.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Bitstream copy	Image
Forensic copy	Mirror image

Imaging

The process of scanning pictures or documents into a computer in order to better manage documents.

Source: RSI, Glossary.

The process of taking an electronic “picture” of a document and storing it on a disk for later retrieval. The stored images cannot be searched, so they are typically linked to records in a database and retrieved when the associated record is located through a database search.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The process of creating a TIFF or PDF image of documents. An image is created that represents how the page would look if the document were printed to paper.

In Re: Actos

In Re: Actos (Pioglitazone) Products Liability Litigation, MDL No. 6:11-md-2299 (W.D. La. July 27, 2012). A product liability action with a Case Management Order (“CMO”) that memorializes the parties’ agreement on a “search methodology proof of concept to evaluate the potential utility of advanced analytics as a Document identification mechanism for the review and production” of Electronically Stored Information. The search protocol provides for the use of a Technology-

Assisted Review tool on the email of four key custodians. The CMO was signed by District Judge Rebecca F. Doherty.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Inaccessible Data

Or "relatively inaccessible data." In contrast with active data, data that has to undergo a restoration process in order to be displayed on computer screen. The two subsets, in order from more accessible to less accessible, are “backup tapes” and “erased, fragmented or damaged data." According to a new line of case law, relatively inaccessible electronic data is the category as to which a court should consider cost-shifting.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Applied Discovery's Glossary, http://www.nysd.uscourts.gov/courtweb/pdf/D02NYSC/03-04265.PDF#page=24

Inactive Record

Inactive records are those Records related to closed, completed, or concluded activities. Inactive Records are no longer routinely referenced, but must be retained in order to fulfill reporting requirements or for purposes of audit or analysis. Inactive records generally reside in a long-term storage format remaining accessible for purposes of business processing only with restrictions on alteration. In some business circumstances, inactive records may be reactivated.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Index

A list of all words in a database (coded or full text) that is used by the software to provide fast access to information. Rather than search the entire database for a word or phrase when a query is built, the software searches the index instead.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Output from a database such as an index to exhibits or documents responsive to a discovery request.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

(n.) In database design, a list of keys (or keywords), each of which identifies a unique record. Indices make it faster to find specific records and to sort records by the index field -- that is, the field used to identify each record.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Webopedia Computer Dictionary, http://www.pcwebopedia.com/TERM/I/FTP.html.

(v.) To create an index for a database, or to find records using an index.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Webopedia Computer Dictionary, http://www.pcwebopedia.com/TERM/I/FTP.html.

Creating a set of rules and data files which define scanned document sets and allow easy and complete retrieval.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A list of Keywords in which each Keyword is accompanied by a list of the Documents (and sometimes the positions within the Documents) where it occurs. Manual indices have been used in books for centuries; automatic indices are used in Information Retrieval systems to identify Documents that contain particular Search Terms.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Index/Coding Field

A database field used to categorize and organize documents. Often user-defined, these fields can be used for searches.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Indexing

Indexing (which is sometimes interchanged with the term "coding") refers to the information that is added to an image to allow it to be found after it is scanned. Objective indexing or coding is one of two types of indexes used in imaging. A template, something like an index card, is attached to the image in the computer and pertinent information is typed into the template, tagging the document for retrieval purposes. Author of the document, box number, date, subject and type of document are all common index fields.

Source: RSI, Glossary.

Universal term for coding and data entry.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

The manual or automatic process of creating an Index. In Electronic Discovery, Indexing typically refers to the automatic constriction of an electronic Index for use in an Information Retrieval System.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

See also:

Bibliographic Coding
Coding
Issue Code
Issue coding
Level coding
Objective coding
Subjective coding
Tag
Taxonomic coding
Verbatim coding

Industrial Internet

The integration of physical machinery (such as generators, refrigerators, industrial equipment) with networked sensors, software, and big data repositories. The goal of the industrial internet is to integrate objects with people, processes, and data. The industrial internet joins the internet of things, machine-to-machine communication and other sources to ingest data from devices, analyze it, and use it to guide the further operation of these systems.

Source: Herb Roitblat, Search 2020: The Glossary.

Industry Standard Architecture (ISA)

ISA (Industry Standard Architecture) is a standard bus (computer interconnection) architecture that is associated with the IBM AT motherboard. It allows 16 bits at a time to flow between the motherboard circuitry and an expansion slot card and its associated device(s).

Source: TechTarget, ISA (Industry Standard Architecture) definition, http://searchwindowsserver.techtarget.com/definition/ISA-Industry-Standard-Architecture

Information Governance

The specification of decision rights and an accountability framework to encourage desirable behavior in the valuation, creation, storage, use, archival and deletion of information. It includes the processes, roles, standards and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals (as defined by Gartner).

Source: IGRM White Paper

Information Governance Reference Model (IGRM)

A framework and responsibility model for cross-functional and executive dialogue that serves as a catalyst for defining a unified governance approach to information by linking business value and legal duties to the information assets.

Source: IGRM White Paper

Information Need

In Information Retrieval, the information being sought in a search or review effort. In E-Discovery, the Information Need is typically to identify Documents responsive to a request for production, or to identify Documents that are subject to privilege or work-product protection.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Information Retrieval

The science of how to find information to meet an Information Need. While modern Information Retrieval relies heavily on computers, the discipline predates the invention of computers.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The process of identifying documents or other records in a corpus that are relevant to the user's interest or information need. Information retrieval is a term of art in computer science that is typically broader than search (a hypernym), but is also frequently used as a synonym for search.

Source: Herb Roitblat, Search 2020: The Glossary.

Input

The transfer of data from keyboard or external storage device to computer memory.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Input Device

Any object which allows a user to communicate with a computer by entering information or issuing commands (e.g. keyboard, mouse or joystick).

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

Input/Output

See: I/O (Input/Output)

Instant Messaging

See: IM (Instant Messaging)

Institute of Electrical and Electronic Engineers

See: IEEE (Institute of Electrical and Electronic Engineers)

Integrated Drive Electronics

See: IDE (Integrated Drive Electronics)

Integrated Services Digital Network (ISDN)

An all digital network which can carry data, video and voice.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Integration

The ability of two systems, whether hardware or software, to interface with one another. Integrated systems are often designed to share data in a way specifically intended to reduce redundant data entry.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Intelligent Character Recognition

See: ICR (Intelligent Character Recognition)

Interface

A mechanical or electrical link connecting two or more pieces of equipment together.

Source: RSI, Glossary.

A point of demarcation between two devices where the electrical signals, connectors, timing and handshaking are defined.

Source: RSI, Glossary.

A connection between any two elements in a computer system.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Interlaced

TV & CRT pictures must constantly be "refreshed". Interlace is to refresh every other line once/refresh cycle. Since only half the information displayed is updated each cycle, interlaced displays are less expensive than "non-interlaced." However, interlaced displays are subject to jitters. The human eye/brain can usually detect displayed images which are completely refreshed at less than 30 times per second.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Internal Inquiry

A close examination of a matter in a search for information or truth that is internal to a company.

Source: RenewData, Glossary (10/5/2005).

Internal Response Curve

From Signal Detection Theory, a tool for estimating the number of Relevant and Non-Relevant Documents in a Population, or the number of Documents that fall above and below a particular Cutoff. The use of Internal Response Curves for this purpose assumes that the scores yielded by a Machine Learning Algorithm for Relevant Documents obey a Gaussian Distribution, and the scores for Non-Relevant documents obey a different Gaussian Distribution. These distributions are then used to predict the number of Relevant and Non-Relevant Documents in any given range of scores.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

International Standards Organization (ISO)

ISO is an independent, non-governmental international organization with a membership of 161 national standards bodies. Through its members, it brings together experts to share knowledge and develop voluntary, consensus-based, market relevant International Standards that support innovation and provide solutions to global challenges.

Source: About ISO, http://www.iso.org/iso/home/about.htm

Internet

The world-wide collection of inter-connected networks that all use the TCP/IP protocols and that evolved from the ARPANET of the late 1960's and early 1970's.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A worldwide computer network containing a broad array of services and information available to any individual with a PC and the paid connection.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

The interconnecting global public network made by connecting smaller shared public networks. The most well-known Internet is the Internet, the worldwide network of networks which use the TCP/IP protocol to facilitate information exchange.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Intranet Extranet

Internet of Everything

See: Internet of Things (IoT)

Internet of Things (IoT)

The emerging idea that every-day things like refrigerators, garage door openers, and thermostats can and should be connected to the internet where they can communicate with one another, be remotely controlled, and do other tasks such as ordering ice cream when the current supply in the freezer is nearly gone. In order to be connected usefully to the internet, things need sensors, network connectivity, software, and processing capability to collect and exchange data.

Source: Herb Roitblat, Search 2020: The Glossary.

Internet Protocol Address (IP Address)

A set of numbers which uniquely identifies an address on a network.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The internationally recognized location of a specific computer or server; used for internet.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A string of four numbers separated by periods used to represent a computer on the Internet.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Internet Publishing

Specialized imaging software that allows large volumes of paper documents to be published on the Internet or intranet. These files can be made available to other departments, offsite colleagues or the public for searching, viewing and printing.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Internet Service Provider (ISP)

A company that provides access to the internet through its own equipment to users and charges a monthly or hourly rate for providing that service.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A business that delivers access to the Internet.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Internetwork Packet Exchange (IPX)

A communications protocol used by Novell networks.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

External links:

Webopedia Computer Dictionary, http://www.webopedia.com/TERM/I/IPX.html

Interpolated / Uninterpolated

Most scanners have a maximum pixel-per-inch resolution before they start guessing or interpolating the data. Interpolated files require the computer to simulate data in an image file, while uninterpolated files hold only data that is accurate to the original. Uninterpolated resolution is, therefore, preferred for accurate scanning.

Source: RSI, Glossary.

Interrogatory

In a civil action, an interrogatory is a list of questions one party sends to another as part of the discovery process. The recipient must answer the questions under oath and according to the case's schedule. Because attorneys may help their clients answer interrogatories, interrogatory responses tend to be more finely crafted than answers to deposition questions. The number of questions included in an interrogatory is usually limited by court rule. For example, under the Federal Rules of Civil Procedure, each party may only ask each other party 25 questions via interrogatory unless the court gives permission to ask more. See Rule 33.

Source: Legal Information Institute, Interrogatory, https://www.law.cornell.edu/wex/interrogatory

See also:

Discovery request Document request Request for admission

Intranet

A private or internal network that uses standard internet protocols so it has the appearance of a Web site.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A network of interconnecting smaller private networks that are isolated from the public Internet.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A computer network usually that usually restricts access only to those within the firm or corporation. The term is often used to discuss documents stored as web pages, but it can be extended to any kind of computer resource. An internal (to the company) version of the internet.

See also:

Internet Extranet

Inverted Index

An index that maps a keyword to the list of documents that contain the keyword.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Investigation

An inquiry usually initiated by a governmental agency.

Source: RenewData, Glossary (10/5/2005).

IoT

See: Internet of Things (IoT)

IP Address

See: Internet Protocol Address (IP Address)

IPC

See: Image Processing Card (IPC)

IPv6

The latest version of the Internet Protocol, version 6 (IPv6), which provides an identification and location system for computers and other devices on networks. IPv4 represented these locations as a string of four numbers: 000.000.000.000, where each number could be between 0 and 256. IPv6 instead creates addresses out of longer digit strings represented as eight groups of four digits (0 to 65535) written in hexadecimal (base 16), with a colon between groups, for example 2001:0db8:0000:0042:0000:8a2e:0370:7334. Many more things can be represented using IPv6 (2128 or 3.4×1038) than were possible under IPv4 (232, approximately 4.3 billion addresses). IPv4 allowed only 4.3 billion devices to be connected directly to the Internet, but many times that would be required when the Internet of Things is fully implemented, for example.

Source: Herb Roitblat, Search 2020: The Glossary.

IPX

See: Internetwork Packet Exchange (IPX)

ISA

See: Industry Standard Architecture (ISA)

ISDN

See: Integrated Services Digital Network (ISDN)

ISIS Scanner Driver

A specialized application used for communication between scanners and computers.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

ISO

See: International Standards Organization (ISO)

ISO 9660 CD Format

The International Standards Organization format for creating CD-ROMs that can be read worldwide.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

ISP

See: Internet Service Provider (ISP)

Issue Code

Term for a code used to designate a case-specific issue. Issue codes are used to maintain consistency, eliminate spelling errors, and speed up search queries.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Issue codes are used to classify documents as being relevant to one or more of the case issues. Case issues are the things about the case that need to proved during the litigation.

See also:

Bibliographic Coding
Coding
Indexing
Issue coding
Level coding
Objective coding
Subjective coding
Tag
Taxonomic coding
Verbatim coding

Issue Code(s) / Issue Coding

One or more subcategories of the overall Information Need to be identified in a search or review effort; the act of generating such subcategories of the overall Information Need. Examples include specification of the reason(s) for a determination of Relevance or Non-Relevance, Coding of particular subcategories of interest, and Coding of privileged, confidential, or significant (“hot”) Documents.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Issue Coding

A process where content is evaluated to determine whether it relates to topics of interest in a lawsuit or similar proceeding and then the results of the evaluation are logged.

See also:

Bibliographic Coding	Level coding	Taxonomic coding
Coding	Objective coding	Verbatim coding
Indexing	Subjective coding	
Issue Code	Tag	

IStorage

The standard for storing additional file information in MS Office files.

Source: Ibis Consulting, Glossary.

Iterative Search

Formal search that includes executing, tracking. reporting and measure impact, and iterate through sets of multiple logical queries. See also, Formal Search.

Source: EDRM Search Glossary.

Iterative Training

The process of repeatedly augmenting the Training Set with additional examples of Coded Documents until the effectiveness of the Machine Learning Algorithm reaches an acceptable level. The additional examples may be identified through Judgmental Sampling, Random Sampling, or by the Machine Learning Algorithm, as in Active Learning.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

J

Jaccard Index

A measure of the consistency between two sets (e.g., Documents Coded as Relevant by two different reviewers). Defined mathematically as the size of the intersection of the two sets, divided by the size of the union (e.g., the number of Documents coded as Relevant by both reviewers, divided by the number of Documents identified as Relevant by one or the other, or both reviewers). It is typically used as a measure of consistency among review efforts, but also

may be used as a measure of similarity between two Documents represented as two Bag of Words. Jaccard Index is also referred to as Overlap or Mutual F1. Empirical studies have shown that expert reviewers commonly achieve Jaccard Index scores of about 50%, and that scores exceeding 60% are rare.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A measure of agreement or efficacy. The Jaccard index compares the number of documents selected as responsive by both assessors divided by the number of documents that are selected as responsive by either assessor. If assessor A identifies 20 documents as responsive and assessor B identifies 25 documents as responsive, and they agree on their identification of 10 documents as responsive, then the numerator would be 10 and the denominator would 20 + 25 – 10, or 10/35 or 28.6%.

Source: Herb Roitblat, Predictive Coding Glossary.

JASIST Study

A 2009 study (Herbert L. Roitblat, Anne Kershaw & Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y. FOR INFO. SCI. & TECH. 70 (2010)), showing that the Positive Agreement between each of two Technology-Assisted Review methods, and a prior production to the Department of Justice, exceeded the Positive Agreement between each of two Manual Review processes and the same production. Also referred to as the EDI Study.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Java

Java is a computer programming language that is designed for use in the multiple-computer environment of the internet. It can be used to make single-machine applications or be distributed among many computers on a network. In addition, Java can be used to create applets for use within a Web page, which allows users to interact directly with the page. Because Java requires no operating system specific extensions, Java applets run on most operating systems. Java is not the same as JavaScript, a Netscape creation that is easier to learn than Java, but lacks some of the speed and portability of Java.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

HTML	SGML	XML
JavaScript	SGML/HyTime	

JavaScript

When referring to a computer processor or CPU, GHz is a clock frequency, also known as a clock rate or clock speed, representing a cycle of time. An oscillator circuit supplies a small amount of electricity to a crystal each second that is measured in KHz, MHz, or GHz. "Hz" is an abbreviation of Hertz, and "K" represents Kilo (thousand), "M" represents Mega (million), and "G" represents Giga (thousand million).

Source: JavaScript, https://developer.mozilla.org/en-US/docs/Web/JavaScript

See also:

HTML	SGML	XML
Java	SGML/HyTime	

Jaz Drive

A Jaz drive is a small, portable hard disk drive used primarily for backing up and archiving personal computer files. The Jaz drive is sold by Iomega Corporation, the same company that developed the Zip drive. Both the Jaz drive and the disks come in two sizes, 1 GB and 2 GB. The two sizes look similar, but a 2 GB disk is not compatible with a 1 GB Jaz drive. The 2 GB Jaz drive can use both disk sizes. Internal and external Jaz drives are available. The Jaz drive uses the Small Computer System Interface (Small Computer System Interface) and requires a SCSI controller.

Source: TechTarget, Jaz Drive definition, http://searchstorage.techtarget.com/definition/Jaz-drive

See also:

Disk drive	Portable drive	Zip drive
Floppy disk drive	Storage device	
Magneto-optical drive	Tape drive	

JMS (Jukebox Management Software)

The most fundamental purpose of jukebox management software is to provide drive letter access to a jukebox. This lets applications and users directly access the jukebox without having to use some exotic programming. Jukeboxes are too complex for the standard computer desktop, so management software simplifies the manner in which users gain access to jukeboxes for storing and retrieving files.

Source: KOM Software, https://www.komsoftware.com/news/news-reviews/jukebox-management-software.html

Joint Photographic Expert Group (JPEG)

One format of electronic graphic image file supported by the web. JPEG files end with the suffix jpg. Other image formats include Graphic Interchange Format (GIF) and Portable Network Graphic (PNG).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A compression algorithm for condensing the size of image files. JPEGs are very helpful in allowing access to full-screen image files on-line because they require less storage and therefore are quicker to download into a web page.

Source: RSI, Glossary.

An image compression format used for storing color photographs and images.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

An image compression standard for photographs.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A standard algorithm for the compression of digital images and the files that result from encoding an image using this algorithm.

See also:

GIF
Graphic Interchange File
Image file format
Multi-page TIFF
PDF
PNG
Portable Document Format
Portable network graphic
Searchable TIFF
Single-page TIFF
TIFF

JOLT Study

A 2011 study (Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011)), available at http://jolt.richmond.edu/v17i3/article11.pdf, that used data from TREC 2009 to show that two Technology-Assisted Review processes (one using Machine Learning and one using a Rule Base) generally achieved better Recall, better Precision, and greater efficiency than the TREC Manual Review process. Also known as the Richmond Journal Study, or the Richmond Study.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

JPEG

See: Joint Photographic Expert Group (JPEG)

Judgmental Sample / Judgmental Sampling

A method in which a Sample of the Document Population is drawn, based at least in part on subjective factors, so as to include the "most interesting" Documents by some criterion; the Sample resulting from such method. Unlike a Random Sample, the statistical properties of a Judgmental Sample may not be extrapolated to the entire Population. However, an individual (such as a Quality Assurance auditor or an adversary) may use Judgmental Sampling to attempt to uncover defects. The failure to identify defects may be taken as evidence (albeit not statistical evidence, and certainly not proof) of the absence of defects.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A sampling process where the objects are selected on the basis of some person's judgments about their relative importance rather than on a random basis. Judgmental sampling sometimes refers to the use of a seed set or preselected documents used to train predictive coding systems. Unlike random samples, judgmental samples are not typically representative of the collection or population from which they are drawn. It is not possible to extrapolate from the characteristics of a judgmental sample to the characteristics of the population or collection.

Source: Herb Roitblat, Predictive Coding Glossary.

Jukebox

Automated disk changer for high-performance, centralized storage for multifunction CD-ROM's & optical disks.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Jukebox Management Software

See: JMS (Jukebox Management Software)

Jump Drive

Also known as keychain drive, thumb drive and USB flash drive.

K

KB (Kilobyte)

One kilobyte of data is equal to one thousand bytes.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The equivalent of 1,000 (actually 1,024) bytes. Indicates (1) size of the storage area on a disk, such as 32KB = 32,768 bytes, or (2) amount of main memory (RAM) in the computer, such as 640K = room to store 640,000 bytes of instructions.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

One thousand bytes of data is 1K of data.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Bit
Byte
MB - megabyte
GB - gigabyte
TB - terabyte
PB - petabyte
EB - exabyte

Kerning

Adjusting the spacing between two letters from the "normal" spacing. Often done to enhance the quality of the typography – for instance in a headline.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Key

A key is a value applied using an algorithm to a string of unencrypted text to produce encrypted text, or vice versa. Key length is a factor in determining the strength of the encryption.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Key Field

A database field used for document searches and retrieval. Synonymous with “index field.”

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Marginalia
Names mentioned in text
Note field
Other number field

Production source | Subject category | Text

Recipient | Summary

Keyboard

The device that allows commands to be typed directly into the computer. Similar to a typewriter keyboard but with special function keys added along the top.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Keychain Drive

See: Jump Drive

Keystroke Monitoring

A form of user surveillance in which the actual character-by-character traffic (that user's keystrokes) are monitored, analyzed, and/or logged for future reference.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Keyword

A specific word used to search a database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Words related to the case or specific issues, designated by the law firm and generally having their own field in the database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Used in bibliographical coding to indicate that each page in a collection must be reviewed for certain important words and wherever they occur the database must reference the page where they occur.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A word (or Search Term) that is used as part of a Query in a Keyword Search.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

See also:

Ad Hoc Search | Adaptive pattern recognition | Associative retrieval | Boolean search

Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Keyword Index / Indexing

Indexing is a process that inventories the total content of a file and builds a searchable electronic index. This index typically maps from a keyword to all the documents that contain the keyword. Search indexes serve to function as tools designed to facilitate and expedite the retrieval of information. Search engines will use both common and proprietary technology to build indexes and service search queries.

Source: EDRM Search Glossary.

A technique that examines the ESI and builds a searchable electronic index. This index typically maps from a keyword to all the documents that contain the keyword.

Source: EDRM Search Guide Glossary.

Keyword Occurrence

Keyword occurrences are the counts of keywords that appear within the entire search results. When a search query involves multiple keywords or when one or more of the queries produces stemming, wildcard or fuzzy-based variations, a complete count of total occurrences for each keyword is useful for evaluating the value of searching using certain keywords. In some instances, the keyword counts both at an aggregate level (totaled over all the variations) as well as counts based on an individual variation level would each be helpful.

Source: EDRM Search Guide Glossary.

Keyword Search

A common search technique that uses query words (“keywords”) and looks for them in ESI, using an index. A keyword search is a basic search technique that involves searching for one or more words within a collection of documents and returns only those documents that contain the search terms entered. The documents returned by the search engine are called the search results. Keywords often form a basic building block for constructing other more complex compound searches. Such compound searches use other search elements such as Boolean logic.

Source: EDRM Search Glossary.

A very common search technique that uses query words ("keywords") and looks for them in ESI, using an index.

Source: EDRM Search Guide Glossary.

A search in which all Documents that contain one or more specific Keywords are returned.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A method of searching for documents that possess keywords specified by a user.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A search using a full text search filter. A client search term list is applied to a full text index to find responsive files.

Source: Ibis Consulting, Glossary.

A search for documents containing one or more words that are specified by a user.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

KHz (Kilohertz)

The kilohertz, abbreviated kHz or KHz, is a unit of alternating current (AC) or electromagnetic (EM) wave frequency equal to one thousand hertz (1,000 Hz). The unit is also used in measurements or statements of signal bandwidth.

Source: TechTarget, kHz (kilohertz) definition, http://searchnetworking.techtarget.com/definition/kHz

See also:

Hz MHz GHz

Kilobyte

See: KB (Kilobyte)

Kilohertz

See: KHz (Kilohertz)

Kleen

Kleen Prods. LLC v. Packaging Corp. of Am., Case No. 1:10-cv-05711, various Pleadings and Tr. (N.D. Ill. 2012). A federal case in which plaintiffs sought to compel defendants to use Content Based Advanced Analytics (CBAA) for their production, after defendants had already employed a complex Boolean Searches to identify Responsive Documents. Defendants advanced Elusion scores of 5%, based on a Judgmental Sample of custodians, to defend the reasonableness the Boolean Search. After two days of evidentiary hearings before (and many conferences with) Magistrate Judge Nan R. Nolan, plaintiffs withdrew their request for CBAA, without prejudice.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

KM (Knowledge Management)

Control of and access to content in all an organization's various databases (CMS, DMS, WP, etc.).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Knowledge Management

See: KM (Knowledge Management)

Kofax Board

The generic term for a series of image processing boards manufactured by Kofax Imaging Processing. These are used between the scanner and the computer, and perform realtime image compression and decompression for faster image viewing, image enhancement, and corrections to the input to account for conditions such as document misalignment, "speckles," etc.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

L

L600 Code Series

A UTBMS Code Set exclusively for e-discovery purposes created by the LEDES Oversight Committee ("LOC") Board.

Source: EDRM Metrics Glossary

LAN (Local Area Network)

A LAN is a group of associated computers which share a common communications line and server within the same geographic area. Typically, LAN users share applications and data storage on the same server.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A system of interconnected computers with a central storage unit (the fileserver), cabling system (the topology), and specific network software (the NOS).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Usually a collection of PC's, connected by cable. Landscape Mode The image is represented on the page or monitor such that the width is greater than the height.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Usually refers to a network of computers in a single building or other discrete location.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A computer network that connects several computers located nearby, allowing them to share files and devices such as printers.

See also:

Client/server network	Peer-to-peer network	WAN - wide area network
MAN - metropolitan area network	SAN - storage area network	
Network	Stand alone computer	

Landscape Orientation

In word processing and desktop publishing, the terms portrait and landscape refer to whether the document is oriented vertically or horizontally. A page with landscape orientation is wider than it is tall.

Source: Webopedia, Landscape Orientation, http://www.webopedia.com/TERM/L/landscape.html

See also:

Portrait orientation

Language Modeling

Computing a model of the relationships among words in a collection. Language modeling is used in speech recognition to predict what the next word will be based on the pattern of preceding words. Language modeling is used in information retrieval and predictive coding to represent the meaning of words in the context of other words in a document or paragraph.

Source: Herb Roitblat, Predictive Coding Glossary.

Laptop Computer

A portable computer, usually weighing less than 15 pounds.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Computer
File server
Microcomputer
Minicomputer
Notebook computer
Personal computer
Workstation

Laser Disc

Same as an optical CD, except 12" in diameter.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

CD
CD-R
CD-ROM
CD-RW
Disc
Disk
Diskette
DVD
DVD-ROM
Floppy disk
Hard disk
Hard drive
Jaz disk
Magnetic disk
Magnetic storage media
Media
Optical disk
Storage media
WORM disk
Zip disk

Latency

The time it takes to read a disk (or jukebox), including the time to physically position the media under the read/write head, seek the correct address and transfer it.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Latent Semantic Analysis (LSA)

(LSA) a statistical method for finding the underlying dimensions of correlated terms. For example, words like law, lawyer, attorney, lawsuit, etc. All share some meaning. The presence of any one of them in a document could be recognized as indicating something consistent about the topic of the document. Latent Semantic Analysis uses statistics to allow the system to exploit these correlations for concept searching and clustering.

Source: Herb Roitblat, Predictive Coding Glossary.

Latent Semantic Indexing (LSI)

The use of latent semantic analysis to index a collection of documents.

Source: Herb Roitblat, Predictive Coding Glossary.

Latent Semantic Indexing / Latent Semantic Analysis

Latent semantic indexing (sometimes also referred to as Latent Semantic Analysis) is a technology that analyzes co-occurrence of keyword terms in the document collection. In textual documents, keywords exhibit polysemy as well as synonymy. Latent Semantic Indexing refers to the additional factor that certain keywords are related to the concept in that they appear together. These relationships can be “is-a” relationship such as “motorcycle is a vehicle” or a containment relationship such as “wheels of a motorcycle”. Support Vector Machines, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, and others.

Source: EDRM Search Glossary.

A Feature Engineering Algorithm that uses linear algebra to group together correlated Features. For example, "Windows, Gates, Ballmer" might be one group, while "Windows, Gates, Doors" might be another. Latent Semantic Indexing underlies many Concept Search tools. While Latent Semantic Indexing is used for Feature Engineering in some Technology-Assisted Review tools, it is not, per se, a Technology-Assisted Review method. Also referred to as Latent Semantic Analysis.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Leading

Pronounced "ledding," the amount of space between lines of printed text.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Legacy Data

Information the development of which an organization may have invested significant resources and has retained its importance, but has been created or stored by the use of software and/or hardware that has been rendered outmoded or obsolete.

Source: Merrill Corporation, Electronic Discovery Glossary.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Information in the development process that may have significant resources invested into it that has been produced and/or stored on software or hardware that has become obsolete.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Information created or stored on software and/or hardware that is outmoded or obsolete.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Legal Hold

A legal hold is a communication issued as a result of current or anticipated litigation, audit, government investigation or other such matter that suspends the normal disposition or processing of records. The specific communication to business or IT organizations may also be called a "hold," "preservation order," "suspension order," "freeze notice," "hold order," or "hold notice."

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Lempel-Zif & Welch (LZW)

A common, lossless compression standard for computer graphics – used for the majority of TIFF files. Typical compression ratios are 4/1.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Level Coding

Used in bibliographic coding to indicate that certain document types will get a more thorough extraction of data than others. Thus they get a deeper "level" of coding.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Bibliographic Coding
Coding
Indexing
Issue Code
Issue coding
Objective coding
Subjective coding
Tag
Taxonomic coding
Verbatim coding

Line Screen

The number of halftone dots that can be printed per inch. As a general rule, newspapers print at 65 to 85 lpi, large city newspapers at 100 or 120 lpi; magazines at 133 or 150 lpi; and, glossy, "coffee table" books at 175 to 200.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Linear Review

A Document-by-Document Manual Review in which the Documents are examined in a prescribed order, typically chronological order.

Source

Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Link Object

An object that specifies and maintains the relationship between a linked object and a link source. See embedded object.

Source: Ibis Consulting, Glossary.

See also:

Bibliographic coding	Link source	Object
Embedded object	Linked object	

Link Source

A data object stored in a separate location from the container and whose data is represented in the container by a linked object.

Source: Ibis Consulting, Glossary.

See also:

Bibliographic coding	Link object	Object
Embedded object	Linked object	

Linked Object

An object that is created in a source file and inserted into a destination file.

Source: Ibis Consulting, Glossary.

See also:

Bibliographic coding	Embedded object	Link object

Link source Object

Listserv

An automatic mailing list to which people may subscribe and then send and receive e-mail messages to and from each other.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Litigation Support

See: ALS (Automated Litigation Support)

Litigation Support Manager

The individual who administers the automated litigation support efforts within a law firm or corporate legal department.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Litigation Support System

One of several types of database which holds both a copy document and information about that document. Most systems will hold the full text of the document and allow searches to be conducted against the text and /or any additional information that might be present in the database. The most sophisticated systems can group documents into categories based on their content and even predict their coding by reference to specimen coding provided by a lawyer.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

Live Machine

A computer that is powered up and actively logged in.

Source: EDRM Collection Standards

Load File

A load file is used to import images or coding (the bibliographic information about a document (e.g., To, From, CC, BCC, and Subject fields within an email) into a database. It sets out links between the records in a database and the document image files to which each record pertains. This is a critical deliverable of any processing, scanning, or coding job. Without a correctly structured load file, documents will not properly link to their respective database records.

Source: EDRM Metrics Glossary

A data file that sets out links between the records in a database and the document image files to which each record pertains. This is a critical deliverable of any scanning and coding job.

Without a correctly structured load file, documents will not properly link to their respective database records.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Commonwealth Legal's Litigation Support Glossary, http://commonwealthlegal.com/resources/glossary.html#l.

A file accompanying output data delivered to a client, containing a log of files and images in a format required by the client's document management system.

Source: Ibis Consulting, Glossary.

A text file with entries for application information and comments. Typically used in automated litigation support to carry instructions about a document image collection for linking to a database program.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A file that relates to a set of scanned images and indicates where individual pages belong together as documents.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A standardised file used for loading metadata and other information into a Litigation Support System. Such files generally contain metadata in a delimited format together with information used to load a copy document into the system. A standardised file used for loading metadata and other information into a Litigation Support System. Such files generally contain metadata in a delimited format together with information used to load a copy document into the system.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

Load File Format

The specific format for load file data, including the load file, CD directory structure, CD content and CD content requirements specific to a particular client or project.

Source: Ibis Consulting, Glossary.

Local Area Network

See: LAN (Local Area Network)

Log

A hard copy record book, usually of entries into a database but also of documents received, documents undergoing quality control, or documents shipped out.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Logical Evidence File

With a logical evidence file, you can selectively choose which files or folders you want to preserve, instead of acquiring the entire drive. Unlike copying files from a device and altering critical metadata, logical evidence files preserve the original files as they existed on the media and include additional information such as file name, file extension, last accessed, file created, last written, entry modified, logical size, physical size, MD5 hash value, permissions, starting extent, and original path of the file.

Source: EDRM Collection Standards

Logical Target

When forensic imaging process targets a logical portion of the media such as the C:\ drive or other logical volume or partition.

Source: EDRM Collection Standards

Logical Unitization

The assembly of individually scanned pages into documents:

- **Physical unitization** utilizes actual objects such as staples, paper clips and folders to determine pages that belong together as documents for archival and retrieval purposes.
- **Logical unitization** is the process of human review of each individual page in an image collection using logical cues to determine pages that belong together as documents. Such cues can be consecutive page numbering, report titles, similar headers and footers and other logical cues.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Logistic Regression

A state-of-the-art Supervised Learning Algorithm that estimates the Probability that a Document is Relevant, based on the Features it contains.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Lookup Table

A predefined set of entries from which a user may pick a name rather than enter the name directly into a database field.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Lossless Compression

Exact construction of image, bit-by-bit, with no loss of resolution or color fidelity.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Lossy Compression

Reduces storage size of image by reducing the resolution and color fidelity while maintaining minimum acceptable standard for general use.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

LSA

See: Latent Semantic Analysis (LSA)

LZW

See: Lempel-Zif & Welch (LZW)

M

Machine Learning

The use of a computer Algorithm to organize or Classify Documents by analyzing their Features. In the context of Technology-Assisted Review, Supervised Learning Algorithms (e.g., Support Vector Machines, Logistic Regression, Nearest Neighbor, and Bayesian Classifiers) are used to infer Relevance or Non-Relevance of Documents based on the Coding of Documents in a Training Set. In Electronic Discovery generally, Unsupervised Learning Algorithms are used for Clustering, Near-Duplicate Detection, and Concept Search.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A process for using computer algorithms and methods to implement a decision, prediction, or categorization process. Machine learning processes typically apply information derived from examples to predict, categorize, or decide about previously unseen objects. Machine learning methods have largely been derived from the science of pattern recognition, brain simulation, learning theory, and decision theory. Machine learning is closely related to statistical modeling.

Source: Herb Roitblat, Search 2020: The Glossary.

A branch of computer science that deals with designing computer programs to extract information from examples. For example, properties that distinguish between responsive and nonresponsive documents may be extracted from example documents in each category. The goal is to predict the correct category for future untagged examples based on the knowledge

extracted from the previously classified examples. Example approaches include neural networks, support vector machines, Bayesian classifiers and others.

Source: Herb Roitblat, Predictive Coding Glossary.

Macro

A pre-programmed keystroke or combination of keystrokes to activate a sequence of instructions.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Magenta

Used in four color printing. Reflects blue & red and absorbs green.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Magnetic Disk Emulation (MDE)

Software that makes a jukebox look and operate like a hard-drive such that it will respond to all the I/O commands ordinarily sent to a hard drive.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Magnetic Ink Character Recognition (MICR)

The process used by banks to encode checks.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Magnetic Storage Media

Includes, but is not limited to, hard drives (also known as "hard disks"), backup tapes, CD-ROMs, DVD-ROMs, Jaz and Zip drives, and floppy discs, all used singly or in combination in, or in conjunction with, your computers and any and all backup and archive systems for the same.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

See also:

Backup
Backup tape
DAT - digital audio tape
Data extraction
Digital audio tape
Disaster recovery tape

DLT - digital linear tape
Magnetic storage media
Media
QIC - quarter inch cartridge
Tape
CD
CD-R
CD-ROM
CD-RW
Disc
Disk
Diskette
DVD
DVD-ROM
Floppy disk
Hard disk
Hard drive
Jaz disk
Laser disc
Magnetic disk
Media
Optical disk
Storage media
WORM disk
Zip disk

Magneto-Optical

A disk storage technology which competes with traditional magnetic hard disks. Form factors are 3.5", 5.25" and 12". Advantages are that one 5.25" magneto-optical drive can store about 1.3GB (3 1/2" hold up to 230MB); media is removable and portable; and, can last for 20 years – ideal for archival storage. The disadvantages are cost, traditionally slower disk access and longer disk write times. The information is written on the disk by changing the polarity with strong magnets and read by a laser by sensing the magnetic flux changes (1's or 0's). This technology is re-usable.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Magneto-Optical Drive

A drive that combines laser and magnetic technology to create high-capacity erasable storage.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Disk drive
Floppy disk drive
Jaz drive
Portable drive
Storage device
Tape drive
Zip drive

Mail

See: E-Mail (Electronic Mail)

Mail Application Program Interface (MAPI)

A Windows software standard that has become a popular email interface used by MS Exchange, GroupWise, and other email packages.

Source: Ibis Consulting, Glossary.

This Windows software standard has become a popular e-mail interface and is used by MS Exchange, GroupWise, and other e-mail packages.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Mail Container

An area in memory or on a storage device where email is placed. In email systems, each user has a private mailbox. When the user receives email, the mail system automatically puts it in the mailbox. The mail system allows you to scan mail that is in your mailbox, copy it to a file, delete it, print it, or forward it to another user. The mailbox format used by Microsoft Exchange email systems is PST, while Lotus Notes uses NSF files.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

A container for electronic mail data (messages and attachments) that contains one or more mail messages. There are multi-mail containers like PST, NSF, Netscape mail containers, etc. and single mail containers like EML and MSG 3.

Source: Ibis Consulting, Glossary.

See also:

Container	NSF	RFC822
EML	OST	Single-mail archive
MSG	PST	Single-mail container
Multi-mail container	RFC compliant email	SMTP

Mailbox

See: Mail Container

MAN (Metropolitan Area Network)

A metropolitan area network (MAN) is a network that interconnects users with computer resources in a geographic area or region larger than that covered by even a large local area network (LAN) but smaller than the area covered by a wide area network (WAN). The term is applied to the interconnection of networks in a city into a single larger network (which may

then also offer efficient connection to a wide area network). It is also used to mean the interconnection of several local area networks by bridging them with backbone lines. The latter usage is also sometimes referred to as a campus network.

Source: TechTarget, metropolitan area network (MAN) definition, http://searchnetworking.techtarget.com/definition/metropolitan-area-network-MAN

See also:

Client/server network
Peer-to-peer network
Stand alone computer
LAN - local area network
SAN - storage area network
WAN - wide area network
Network

Management Information Systems (MIS)

A management information system (MIS) focuses on the management of information systems to provide efficiency and effectiveness of strategic decision making. The concept may include systems termed transaction processing system, decision support system, expert system, or executive information system. The term is often used in the academic study of businesses and has connections with other areas, such as information systems, information technology, informatics, e-commerce and computer science; as a result, the term is used interchangeably with some of these areas.

Management information systems (plural) as an academic discipline studies people, technology, organizations, and the relationships among them. This definition relates specifically to "MIS" as a course of study in business schools. Many business schools (or colleges of business administration within universities) have an MIS department, alongside departments of accounting, finance, management, marketing, and may award degrees (at undergraduate, master, and doctoral levels) in Management Information Systems.

Source: Wikipedia, Management information system, https://en.wikipedia.org/wiki/Management_information_system

Manual Review

The practice of having human reviewers individually read and Code the Documents in a Collection for Responsiveness, particular issues, privilege, and/or confidentiality.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Map-Reduce

A computational pattern in which complex computations are broken down into two kinds of steps. In the Map step, the data are processed in parallel, typically on a large number of processors. The results of the Map step are then combined in the Reduce step to yield a final result. The Map-Reduce pattern is typically used on top of Hadoop to process big data.

Source: Herb Roitblat, Search 2020: The Glossary.

MAPI

See: Mail Application Program Interface (MAPI)

MAPI Mail Near-Line

Documents stored on optical disks or compact disks that are housed in the jukebox or CD changer and can be retrieved without human intervention.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Margin of Error

The maximum amount by which a Point Estimate might likely deviate from the true value, typically expressed as "plus or minus" a percentage, with a particular Confidence Level. For example, one might express a Statistical Estimate as "30% of the Documents in the Population are Relevant, plus or minus 3%, with 95% confidence." This means that the Point Estimate is 30%, the Margin of Error is 3%, the Confidence Interval is 27% to 33%, and the Confidence Level is 95%. Using Gaussian Estimation, the Margin of Error is one-half of the size of the Confidence Interval. It is important to note that when the Margin of Error is expressed as a percentage, it refers to a percentage of the Population, not to a percentage of the Point Estimate. In the current example, if there are one million Documents in the Document Population, the Statistical Estimate may be restated as "300,000 Documents in the Population are Relevant, plus or minus 30,000 Documents, with 95% confidence"; or, alternatively, "between 270,000 and 330,000 Documents in the Population are Relevant, with 95% confidence." The Margin of Error is commonly misconstrued to be a percentage of the Point Estimate. However, it would be incorrect to interpret the Confidence Interval in this example to mean that "300,000 Documents in the Population are Relevant, plus or minus 9,000 Documents." The fact that a Margin of Error of "plus or minus 3%" has been achieved is not, by itself, evidence of a precise Statistical Estimate when the Prevalence of Relevant Documents is low.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The likely range in which the true population value will be found.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

Confidence Interval

Marginalia

A data field recording the existence of handwriting in the margins of a document.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Handwritten notes in the margin of the page in documents.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Summary
Text

Mastering

Making many copies of a CD-ROM from a single master.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

MB (Megabyte)

One megabyte of data is equal to one million bytes.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The equivalent of 1,000,000 bytes or 700 double-spaced pages of typed material, each page holding approximately 1,500 characters (bytes).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A million bytes of data is a megabyte, or simply a meg.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

1,000 kilobytes (KB) or 1,048,576 bytes.

See also:

Bit
Byte
KB - kilobyte
GB - gigabyte
TB - terabyte
PB - petabyte

EB - exabyte

Mbox

mbox is a common format for storing email messages. An mbox is a single file containing zero or more email messages.

Source: http://www.qmail.org/qmail-manual-html/man5/mbox.html.

MCA (Micro Channel Architecture)

An IBM bus standard.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

MD5

MD5 is an algorithm that is used to verify data integrity through the creation of a 128-bit message digest from data input (which may be a message of any length) that is claimed to be as unique to that specific data as a fingerprint is to the specific individual. The hash value of a file is the unique, identifying number calculated by MD5.

Source: Ibis Consulting, Glossary.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing SearchSecurity.com, http://searchsecurity.techtarget.com/sDefinition/0,,sid14_gci527453,00.html.

See also:

Hash

Hash value

Hashing / Hash / Hash Value

SHA-1

MD5-Known Filter

A filter option that allows for excluding known, commercially available files (such as executable files or commercial software).

Source: Ibis Consulting, Glossary.

See also:

Date filter

Extensions/sizes filter

Filter

Sender/recipient filter

MDE

See: Magnetic Disk Emulation (MDE)

Mean Time Between Failure (MTBF)

Average time between failures. Used to compute the reliability of devices/equipment.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Mean Time To Repair (MTTR)

Average time to repair. The higher the number, the more costly and difficult to fix.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Measure of Success

Reach comfort level that reasonable steps were taken to find document(s), allowing for reasonable determination that document does not exist.

Source: EDRM Search Glossary.

Measurement Bias

Measurement Bias occurs when the act of sampling causes the measurement to be impacted. In e-discovery, measurement bias could occur if the content of the sample is known before the sampling is done. For example, if one were to sample for responsive documents and during the sampling stage, content is reviewed, there is potential for higher-level litigation strategy to impact the responsive documents. If a project manager has communicated the cost of reviewing responsive documents, and it is understood that responsive documents should somehow be as small as possible, that could impact your sample selection. To overcome this, the person implementing the sample selection should not be provided access to the content.

Source: EDRM Search Glossary.

Media

The physical material used to store electronic data. Media includes hard drives, backup tapes, computer disks, CDs, DVDs, PDAs, memory, etc.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Any external data store format, such as CDs, Jaz drives, DLT tapes, DVDs, or diskettes received from clients containing source data.

Source: Ibis Consulting, Glossary.

The material (disk drive, tape, floppy disk, paper, etc.) on which electronic documents have been recorded.

See also:

Backup	Tape	Hard disk
Backup tape	CD	Hard drive
DAT - digital audio tape	CD-R	Jaz disk
Data extraction	CD-ROM	Laser disc
Digital audio tape	CD-RW	Magnetic disk
Disaster recovery tape	Disc	Magnetic storage media
DLT - digital linear tape	Disk	Optical disk
Magnetic storage media	Diskette	Storage media
Media	DVD	WORM disk
QIC - quarter inch cartridge	DVD-ROM	Zip disk
	Floppy disk	

Media Conversion

Moving data from one type of media to another such as tape to CD.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Megabyte

See: MB (Megabyte)

Megahertz (MHz)

A unit of electrical frequency equal to a million cycles per second.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Hz	KHz	GHz

Memory

Internal storage areas in the computer. The term memory identifies data storage that comes in the form of chips, and the word storage is used for memory that exists on tapes or disks. Moreover, the term memory is usually used as a shorthand for physical memory, which refers to the actual chips capable of holding data. Some computers also use virtual memory, which expands physical memory onto a hard disk. See the definitions for two types of physical memory: RAM and ROM.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

Internal storage areas in the computer. The term memory identifies data storage that comes in the form of chips, and the word storage is used for memory that exists on tapes or disks. Moreover, the term memory is usually used as a shorthand for physical memory, which refers to the actual chips capable of holding data. Some computers also use virtual memory, which expands physical memory onto a hard disk. See the definitions for two types of physical memory: RAM and ROM.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

DRAM RAM ROM

Merge

The process of merging various e-mail files (i.e. Microsoft Outlook's .pst) into one file for de-duplication purposes.

Source: RenewData, Glossary (10/5/2005).

To combine data from two separate databases into one.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Meta Tag

An element of HTML that often describes the contents of a Web page, and is placed near the beginning of the page's source code. Search engines use information provided in a meta tag to index pages by subject.

Metadata

The term metadata refers to "data about data". The term is ambiguous, as it is used for two fundamentally different concepts (types). Structural metadata is about the design and specification of data structures and is more properly called "data about the containers of data"; descriptive metadata, on the other hand, is about individual instances of application data, the data content. In this case, a useful description would be "data about data content" or "content about content" thus metacontent.

Source: http://en.wikipedia.org/wiki/Metadata

Data about data. Metadata captures data elements or attributes (name, size, date, type, etc.), data about records or data structures (length, fields, columns, etc.) and data about data (where it is located, how it is associated, ownership, etc.).

Source: RenewData, Glossary (10/5/2005).

Metadata is information about a particular data set which describes how, when and by whom it was collected, created, accessed, modified and how it is formatted. Some metadata, such as file dates and sizes, can easily be seen by users; other metadata can be hidden or embedded and unavailable to computer users who are not technically adept. Metadata is generally not reproduced in full form when a document is printed. (Typically referred to by the not highly informative “short hand” phrase “data about data,” describing the content, quality, condition, history, and other characteristics of the data.)

Source: Merrill Corporation, Electronic Discovery Glossary.

Metadata is information about a particular data set which may describe, for example, how, when, and by whom it was received, created, accessed, and/or modified and how it is formatted. Some metadata, such as file dates and sizes, can easily be seen by users; other metadata can be hidden or embedded and unavailable to computer users who are not technically adept. Metadata is generally not reproduced in full form when a document is printed. (Typically referred to by the less informative shorthand phrase “data about data,” it describes the content, quality, condition, history, and other characteristics of the data.)

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Data that describes how, when and by whom a particular set of data was created, edited, formatted, and processed. Access to meta-data provides important evidence, such as blind copy (bcc) recipients, the date a file or email message was created and/or modified, and other similar information. Such information is lost when an electronic document is converted to paper form for production.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Applied Discovery's Glossary, http://www.lexisnexis.com/applieddiscovery/clientResources/glossary_M.asp

A description or definition of electronic data, or data about data. Often, metadata can only be assessed in certain viewing modes. Metadata can include descriptive HTML tags and information about when a document was created, and what changes have been made on that document.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Data about data. In data processing, metadata provides information about a document or other data managed within an application or environment. There are five types of metadata: file system, document, email, vendor-added, and customer-added.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Data about data. In data processing, metadata provides information about a document or other data managed within an application or environment. There are five types of metadata: file system, document, email, vendor-added, and customer-added. Traditionally, the OCR base was the only data extracted from the documents. With e-discovery, the metadata can also be

obtained. OCR base in the information that is culled from the images contained within each page. (Read: Whatever text is displayed in the image). Contrasting this is the metadata. The metadata is the "foot print" of the document: it the user to review information obtained about the actual document rather than the content.

Source: RSI, Glossary.

In data processing, metadata is data that provides information about or documentation of other data managed within an application or environment. There are two types of metadata: general and file-specific metadata. Metadata is available for any particular Microsoft file in Windows by right-clicking on a file and viewing file Properties. In the Summary tab the Advanced option brings up the list of all possible metadata for that file. See also general metadata and file-specific metadata.

Source: Ibis Consulting, Glossary.

Information about data which describes how, when, and by whom it was received, created, accessed, and/or modified and how it is formatted. Some metadata is visible such as file size and date of creation; most is not visible even when the document is printed.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The data that is attached to files in a computerized filing system. For instance, in a word processing document, the metadata includes: the author, date created, person and date editing the document, the name of the document, the location stored on a hard drive, how many times and when it has been accessed, changed or altered, etc.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Data about a file itself, such as when it was created, modified, and which computer user authored it. For emails this could also include: bcc, date received, opened status, undeliverable, etc. Different metadata are available for different types of electronic files. Metadata can be useful to understanding more about the document and its relevance to the case.

Properties of an electronic file, some of which will be internal and some external, not all of which are necessarily visible when viewing that file.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

See also:

Customer-added metadata	Extrinsic data	File-specific metadata
Document metadata	File parameters	General metadata
Email metadata	File system metadata	Vendor-added metadata

Metadata Search

Metadata search allows searching to be constrained based on certain metadata elements of a document. A general search specification allows for naming the metadata fields, specifying the inherent type of that metadata, and the value to search for.

Source: EDRM Search Glossary.

Metrics DB: Container Files

Container files store one or more files in a compressed form (e.g. RAR or ZIP format).

Source: EDRM Metrics Glossary

Metrics DB: Culling Methods

Procedures used to select a particular set of documents from a larger corpus based on specifically-define criteria. Culling methods are most commonly used as a means to eliminate non-responsive material from a document collection in order to narrow the scope of potentially responsive materials requiring attorney review. Common culling methods include custodial culling, data source culling, date culling, file type culling, domain culling, keyword culling, and deduplication.

Source: EDRM Metrics Glossary

Metrics DB: Custodial Culling

A culling method by which specific data is either selected or removed from a larger set solely based on whether said data is stored and/or maintained by a particular individual on a repository within their administrative control.

Source: EDRM Metrics Glossary

Metrics DB: Data Source Culling

A culling method by which specific data is either selected or removed from a larger set solely based on whether said data originates from or is stored within a particular repository or on certain media.

Source: EDRM Metrics Glossary

Metrics DB: Data Volume Post-Culling

The volume of data remaining after specific culling methods have been applied to a larger data set.

Source: EDRM Metrics Glossary

Metrics DB: Data Volume Post-Deduplication

The amount of data remaining in a data set after duplicate files have been removed.

Source: EDRM Metrics Glossary

Metrics DB: Data Volume Post-Processing

The amount of data in a data set after the extraction of data from contained files and resultant expansion, the application of culling filters and other data reduction methodologies, text extraction and optical character recognition, and other manipulation of native data.

Source: EDRM Metrics Glossary

Metrics DB: Data Volume Pre-Processing

The amount of data collected from various sources prior to the application of any culling methodologies or manipulation or conversion of files in their native format.

Source: EDRM Metrics Glossary

Metrics DB: Data Volume Produced

The total amount of data either delivered to or received from a third party in the context of a legal proceeding. Typically, any amount of data delivered to an opposing or third party has been reviewed for responsiveness, confidentiality and privilege as a precondition of production.

Source: EDRM Metrics Glossary

Metrics DB: Data Volume Reviewed

The total amount of data examined by counsel and classified as responsive to certain claims or issues, as attorney-client communication or privileged work product, confidential, or some other designation prior to it being produced to any third party. A reviewed data volume may also include data that has been classified through the use of advanced analytics technologies according to a defined assisted review process.

Source: EDRM Metrics Glossary

Metrics DB: Date Culling

A culling method by which specific data is either selected or removed from a larger set solely based on criteria related to date or time. Such criteria include the specific date on which a document was created, modified, or accessed, or, the case of email, the date on which a message was sent or received. Typically, a date culling methodology leverages a range of dates within which documents match specific criteria.

Source: EDRM Metrics Glossary

Metrics DB: Dedupe Method

(1) Global/Case. (2) Custodian.

Source: EDRM Metrics Glossary

Metrics DB: Deduplication

In the context of e-discovery, deduplication refers to the reduction of duplicate files based on identical file finger prints or a combination of file finger prints and metadata attributes. Deduplication is used in legal review to reduce the amount of data required for review. Exact duplicates are identified by comparing the hash values of two or more documents. A hash value is a unique identifier associated with a particular document generated by a specific mathematical algorithm based on a document's content and attributes. MD5, SHA-1, and SHA-180 are examples of different hashing algorithms. Deduplication is applied to data sets in different ways including globally (i.e. to an entire data set across custodians – often referred to as "horizontal deduplication"), by custodian (i.e. within each custodian's documents – often referred to as "vertical deduplication"). A near duplicate is a document that is materially similar to another but is different on a bit-level. It is important to note that near deduplication is based on the content of a document, not the hash value, and can be impacted but such factors as standard language, document headers and footers (e.g. email signatures or disclaimers), and OCR quality. Finally, the definition of deduplication within the context of e-discovery is slightly different than that used within data storage management. Storage management often leverages deduplication to store a single instance of a file.

Source: EDRM Metrics Glossary

Metrics DB: Family Count Post-Culling

Represents the number of parent or family files that remain after various filtering and culling methods have been applied.

Source: EDRM Metrics Glossary

Metrics DB: Family Count Post-Deduplication

Represents the number of parent or family files remaining after deduplication process has been applied.

Source: EDRM Metrics Glossary

Metrics DB: Family Count Post-Processing

Typically, this is the first step in preparing data for further filtering and culling. Refers to the total pre-deduplication count of parent or family files/documents after containers and embeddings are extracted. Does not include the original container files (e.g. .zip, .rar, .jar, .tar,etc.), in the file count.

Source: EDRM Metrics Glossary

Metrics DB: Family Count Produced

Total count of parent or family files produced.

Source: EDRM Metrics Glossary

Metrics DB: Family Count Reviewed

The total count of parent or family files that have been reviewed.

Source: EDRM Metrics Glossary

Metrics DB: File Count Pre-Processing

Represents the number of files going into processing. This process will typically count the number of container files but not number of files extracted from containers.

Source: EDRM Metrics Glossary

Metrics DB: File Type Culling

Refers to a processing method by which specific data is either selected or removed from a larger set solely based on the format. Though it can be manipulated, a user can often identify the kind of data stored in a file through the filename extension (e.g. .pdf, .ppt, .docx).

Source: EDRM Metrics Glossary

Metrics DB: Individual Count Post-Culling

Represents the number of files including parents and children remaining after culling methods have been applied.

Source: EDRM Metrics Glossary

Metrics DB: Individual Count Produced

Total count of all files produced including parents and children, counted separately.

Source: EDRM Metrics Glossary

Metrics DB: Individual Count Reviewed

Total count of all files reviewed including parents and children, counted separately.

Source: EDRM Metrics Glossary

Metrics DB: Other Culling

Any other culling method by which certain criteria is used to either select or remove certain data from a larger corpus. Examples include: date range, file type or custodian/source.

Source: EDRM Metrics Glossary

Metrics DB: Threading Culling

An email thread is a file that contains an original email along with the subsequent replies to and/or forwards of that original email. Threading culling allows users to review all of the individual replies and forwarded messages relating to an original email as one inclusive record

or grouped set of records. Users can review emails according to conversations as opposed to viewing fragmented and duplicative emails messages contained within a thread in isolation.

Source: EDRM Metrics Glossary

Metrics DB: Total Hours Reviewed

The total number of hours spent on review by all reviewers (combined). This should include both attorney and litigation support team (paralegal, review specialist, etc) hours. This should not include time spent on processing or culling the data in preparation for review, or the establishment of batches or groups of records for review.

Source: EDRM Metrics Glossary

Metropolitan Area Network

See: MAN (Metropolitan Area Network)

MHz

See: Megahertz (MHz)

MICR

See: Magnetic Ink Character Recognition (MICR)

Micro Channel Architecture

See: MCA (Micro Channel Architecture)

Microcomputer

The next level of computer after the PC, the minicomputer is designed to operate in a multi-user environment. “Mini’s” often use several computer processors in combination.

See also:

Computer	Minicomputer	Workstation
File server	Notebook computer	
Laptop computer	Personal computer	

Microfiche

Reduced sized document(s) filed on sheet microfilm (4" by 6"), containing reduced images of 270 pages or more in a grid pattern. Usually with a human-readable title.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Microfilm

Film on which documents etc. are photographically greatly reduced in size.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Microprocessor

A computer processor on one chip.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The CPU of a PC, the most prevalent of which is the Intel chip (286, 386, 486, and Pentium).

Microsoft Disk Operating System (MS-DOS)

Acronym for disk operating system. The term DOS can refer to any operating system, but it is most often used as a shorthand for MS-DOS (Microsoft disk operating system). Originally developed by Microsoft for IBM, MS-DOS was the standard operating system for IBM-compatible personal computers.

Source: http://www.webopedia.com/TERM/D/DOS.html

Microsoft's disk operating system; used in PC's as the control system.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

DOS

Linux

Microsoft Windows

Network operating system

NOS

Operating system

OS

UNIX

Windows

Xenix

Microsoft DOS

See: Microsoft Disk Operating System (MS-DOS)

Microsoft Windows

A software product that provides an operating environment that runs under MS-DOS, using a GUI that can run different programs at the same time in different windows.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

DOS	NOS	Windows
Linux	Operating system	Xenix
Microsoft DOS	OS	
Network operating system	UNIX	

Migrated Data

Migrated data is information that has been moved from one database or format to another, usually as a result of a change from one hardware or software technology to another.

Source: Merrill Corporation, Electronic Discovery Glossary.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Data that has been transferred from one database or format to another that is generally done when migrated from one form of hardware or software technology to another.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Information that has been moved from one database or format to another.

Migration

The process of moving a computer system and/or its components from one operating environment to another operating environment. Migration also refers to moving data from one storage medium or device to another, as in hardware and software upgrades.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

MIME (Multipurpose Internet Mail Extensions)

A standard for encoding attachments in mail messages. Files with MIME can have numerous, unessential pages.

Source: Ibis Consulting, Glossary.

The unique identifier used to describe which file type is conveyed across a MIME-based protocol such as MIME e-mail or HTTP. The MIME type, contained in certain fields of an email, indicates what kind of computer file is attached to the email so that the system knows how to open the file or otherwise process it. Mime types names conform to an international standard. Registration of MIME types is explained in RFC 2048.

Minicomputer

The next level of computer after the PC, the minicomputer is designed to operate in a multi-user environment. “Mini’s” often use several computer processors in combination.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Computer	Microcomputer	Workstation
File server	Notebook computer	
Laptop computer	Personal computer	

Mirror Image

Used in computer forensic investigations and some electronic discovery investigations, a mirror image is a bit-by-bit copy of a computer hard drive that ensures the operating system is not altered during the forensic examination. May also be referred to as “disc mirroring,” or as a “forensic copy.”

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Bitstream copy	Image
Forensic copy	Imaged copy

Mirroring

Duplication of data for purposes of backup or data distribution.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Computer evidence	Discovery	Electronic evidence
Computer forensics	Electronic discovery / e-discovery	Forensic analysis
Computer investigations		Forensics

MIS

See: Management Information Systems (MIS)

Miss / Missed

A Relevant Document that is not identified as Relevant by a search or review effort. Also referred to as a False Negative.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Miss Rate

The fraction (or Proportion) of truly Relevant Documents that are not identified as Relevant by a search or review effort. Miss Rate = 100% – Recall. Also referred to as the False Negative Rate.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Modem (Modulator-Demodulator)

A device that modulates digital signals to allow their transmission over analog communication facilities. Typically used to allow two computers to communicate over phone lines.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A device which can take digital data from a computer, translate it into analog signals (tones) and transmit the information over telephones lines. Another modem at the receiving computer will receive the information, translate it back from analog to digital and store it. Typical speeds are from 1,200 to 14,400 bits per second. Some modems also correct any errors which occur in the transmission process.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A piece of hardware that lets a computer talk to another computer over a phone line.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

External links:

Webopedia Computer Dictionary, http://www.webopedia.com/TERM/M/modem.html

Modulator-Demodulator

See: Modem (Modulator-Demodulator)

Monitor

A dedicated device that plugs into a graphics board and then displays computer-generated information.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The screen that displays data from the computer. Monitors may be monochrome or color. On notebook computers, they may also be “backlit” or “gas Plasma."

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Monochrome

A display capable of only two colors, usually black & white.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Mosaic

A program used for finding and reading documents on the World-Wide-Web.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Motherboard

The main board into which printed circuit boards or cards are attached to the microprocessor.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Mount

The process of making off-line data available for on-line processing. For example, placing a magnetic tape in a drive and setting up the software to recognize or read that tape. The terms “load” and “loading” are often used in conjunction with, or synonymously with, “mount” and “mounting” (as in “mount and load a tape”). “Load” may also refer to the process of transferring data from mounted media to another media or to an on-line system.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Mouse

A hand-held device that is rolled on the desktop and controls the cursor position on the monitor. Commonly used with software that has a graphical user interface.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

MPEG

MPEG-1 and MPEG-2 are two different standards for full motion video to digital compression/decompression techniques advanced by the Moving Pictures Experts Group. MPEG-1 compresses the bandwidth needed for 30 frames/second of full-motion video (several hundred megabytes) down to about 1.5 Mbits/sec. MPEG-2 only compresses to about 3 Mbits and provides for better image quality when comparing compressed files of the same size. This industry application competes with other compression techniques, know as JPEG, Captain Crunch, Cinepak and Indeo.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

MS-DOS

See: Microsoft Disk Operating System (MS-DOS)

MSG

The Microsoft Outlook Item (.msg) File Format is used to format a Message object, such as an e-mail message, an appointment, a contact, a task, and so on, for storage in the file system.

Source: [MS-OXMSG]: Outlook Item (.msg) File Format- Introduction, http://msdn.microsoft.com/en-us/library/ee160779(v=exchg.80).aspx.

The file format of stand-alone, single-mail message container not contained in multi-mail containers.

Source: Ibis Consulting, Glossary.

Message file. Typically contains an email message.

See also:

Container	NSF	Single-mail archive
EML	OST	Single-mail container
Mail container	PST	SMTP
Mailbox	RFC compliant email	
Multi-mail container	RFC822	

MTBF

See: Mean Time Between Failure (MTBF)

MTTR

See: Mean Time To Repair (MTTR)

Multi-Mail Container

An aggregation of e-mail messages and attachments saved within containers (for example, .PST or .NSF).

Source: Ibis Consulting, Glossary.

See also:

Container	Mail container	MSG
EML	Mailbox	NSF

OST	RFC822	SMTP
PST	Single-mail archive	
RFC compliant email	Single-mail container	

Multi-Page Text

Extracted, multi-page text files, with or without page break characters.

Multi-Page TIFF

Multi-page TIFF images (a single TIFF file with multiple pages). The Bates number name assigned to each of these TIFF files is the Bates number of the first page of the file.

Source: Ibis Consulting, Glossary.

A .tif file comprised of all of the pages contained in the underlying electronic file or hardcopy document prior to its conversion to or scanning into .tif format. As distinguished from the situation where each page of an underlying multi-page document becomes a separate .tif file.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005).

See also:

GIF	JPEG	Portable network graphic
Graphic Interchange File	PDF	Searchable TIFF
Image file format	PNG	Single-page TIFF
Joint photographic expert group	Portable Document Format	TIFF

Multi-Task

The ability to access more than one software application at a time.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The capability to carry out multiple tasks at the same time.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Multi-Thread

Multi-tasking within the same application at the same time.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Multi-User

The capability to have more than one person using a computer system at the same time. A multi-user system allows the sharing of data and peripheral equipment among all users.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Multipurpose Internet Mail Extensions

See: MIME (Multipurpose Internet Mail Extensions)

Multisynch

Analog video monitors which can receive a wide range of display resolutions, usually including TV (NTSC). Color analog monitors accept separate red, green & blue (RGB) signals.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

N

N-Gram

N consecutive words or characters treated as a Feature. In the phrase, "To be or not to be," a word Bigram (i.e., 2-gram) would be "to be"; a word An N-Gram where N = 3 (i.e., a 3-gram). (i.e., 3-gram) would be "to be or"; a Quad-Gram (i.e., 4-gram) would be "to be or not"; and so on. *See also Shingling*.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Names Mentioned in Text

A data field used to classify names that appear in a document other than as the author, recipient, or recipient of a carbon copy.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Note field

Other number field	Recipient	Summary
Production source	Subject category	Text

Native Application

Any application used to create and view a particular application file type.

Source: Ibis Consulting, Glossary.

Native Environment

The original configuration (software, passwords, server configuration, etc.) of a backup tape or e-mail system (i.e. Microsoft Exchange).

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Native File

Application files in their original file format. Also used in the context of delivering native file processing.

Source: Ibis Consulting, Glossary.

A file saved in the format of the original application used to create the file. Dealing with native files can minimize expensive per-page costs for the traditional TIFF and/or PDF processing and will maximize the relevant information available from the file.

Source: RenewData, Glossary (10/5/2005).

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing RenewData's Electronic Evidence Reference Chart, http://www.renewdata.com/wall-chart-signup.html

The source document, as collected from the source computer or server, before any conversion or processing of the document.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

A document produced in the format in which it was originally created.

Native Format

Electronic documents have an associated file structure defined by the original creating application. This file structure is referred to as the “native format” of the document. Because viewing or searching documents in the native format may require the original application (i.e., viewing a Microsoft Word document may require the Microsoft Word application), documents

are often converted to a standard file format (i.e., tiff) as part of electronic document processing.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Native Production

Producing files in the format they were created and maintained is known as a native production.

Source: EDRM Production Guide.

A document produced in the format in which it was originally created.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Natural Language Search

A non-Boolean retrieval method, which, instead of using "and/or" connectors, prepares the search request in ordinary language and is automatically convened by the computer into algorithms.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Naïve Bayes

A Supervised Learning Algorithm in which the relative frequency of words (or other Features) in Relevant and Non-Relevant Training Examples is used to estimate the likelihood that a new Document containing those words (or other Features) is Relevant. Naïve Bayes relies on the simplistic assumption that the words in a Document occur with independent Probabilities, with the consequence that it tends to yield extremely low or extremely high estimates.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Naïve Bayesian Classifier

A system that examines the probability that each word in a new document came from the word distribution derived from trained responsive documents or from trained non-responsive documents. The system is naïve in the sense that it assumes that all words are independent of one another.

Source: Herb Roitblat, Predictive Coding Glossary.

NDLON

National Day Laborer Organizing Network v. U.S. Immigration and Customs Enforcement Agency, Case No. 10-Civ-3488 (SAS), 2012 WL 2878130 (S.D.N.Y. July 13, 2012), a Freedom of Information Act (FOIA) case in which District Judge Shira A. Scheindlin held that "most custodians cannot be 'trusted' to run effective searches because designing legally sufficient electronic searches in the discovery or FOIA contexts is not part of their daily responsibilities," and stated (in dicta) that "beyond the use of keyword search, parties can (and frequently should) rely on latent semantic indexing, statistical probability models, and machine learning to find responsive documents. Through iterative learning, these methods (known as 'computer-assisted' or 'predictive' coding) allow humans to teach computers what documents are and are not responsive to a particular FOIA or discovery request and they can significantly increase the effectiveness and efficiency of searches."

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Near-Duplicate Detection

An industry-specific term generally used to describe a method of grouping together "nearly identical" Documents. Near-Duplicate Detection is a variant of Clustering in which the similarity among Documents in the same group is very strong. It is typically used to reduce review costs, and to ensure consistent Coding. Also referred to as Near-Deduplication.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Nearest Neighbor

A Supervised Learning Algorithm in which a new Document is Classified by finding the most similar Document in the Training Set, and assuming that the correct Coding for the new Document is the same as the most similar one in the Training Set.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A statistical procedure that classifies objects, such as documents, according to the most similar item that has already been assigned a category label. This approach uses a set of labeled examples to classify subsequent unlabeled items, by choosing the category assigned to the most similar labeled example (its nearest neighbor) or examples. K-nearest neighbor classification uses the k most similar classified objects to determine the classification of an unknown object.

Source: Herb Roitblat, Predictive Coding Glossary.

Negative Predictive Value (NPV)

The fraction (Proportion) of Documents that are identified as Non-Relevant by a search or review effort, that are in fact Non-Relevant. The complement of Precision; that is, Negative Predictive Value is computed the same way as Precision when the definitions of Relevant and Non-Relevant are transposed.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Nesting

Document nesting occurs when one document is inserted within another document (i.e., an attachment is nested within an email; graphics files are nested within a Microsoft Word document).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

NetWare Loadable Module (NLM)

An application that runs as part of the network operating system (NOS) of a Novell NetWare server.

Source; Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Network

A group of connected computers that allow people to share information and equipment (e.g. local area network (LAN), wide area network (WAN), metropolitan area network (MAN), storage area network (SAN), peer-to-peer network, client-server network).

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

Multiple computers connected together so that they function as a multi-user system. A network may be a local area network (LAN) or a wide area network (WAN).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A group of computers or devices that is connected together for the exchange of data and sharing of resources.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Client/server network
Peer-to-peer network
WAN - wide area network
LAN - local area network
SAN - storage area network
MAN - metropolitan area network
Stand alone computer

Network Interface Card (NIC)

The card inside a computer that enables the establishment of a network connection.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Network Operating System (NOS)

Software which directs the overall activity of networked computers.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

The operating system that supports network operations.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

DOS
Microsoft Windows
UNIX
Linux
Operating system
Windows
Microsoft DOS
OS
Xenix

Network Topology

The wiring, connections, and adapter boards that interconnect computers on a network. The three standard topologies for PCs are Ethernet, IBM Token Ring, and ARCnet.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Neural Network

An approach to machine learning where the elements of the process resemble simulated neurons. Neurons, in turn, are thought to be the primary computational elements in the brain. Each element in a neural network receives some set of inputs, either from the environment or from other neurons. It then computes an output based on its inputs. Networks of these elements are capable of quite sophisticated computations. Computing with neural networks is also called brain-style computation.

Source: Herb Roitblat, Search 2020: The Glossary.

NIC

See: Network Interface Card (NIC)

NLM

See: NetWare Loadable Module (NLM)

Node

Any device connected to network. PCs, servers, and printers are all nodes on the network.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Noise Word Filter

To avoid creating an overly inclusive index, most indices utilize a noise word filter. Noise word filters includes a customized list of terms that are overlooked or ignored during indexing. Some common noise words include ‘a’, ‘and’, ‘the’, ‘from’, and ‘because’.

Source: EDRM Search Glossary.

Non-Interlaced

When each line of the video image is scanned separately. Computer monitors use non-interlaced video.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Non-Mail Data

Extractable, standard and (some non-standard) mail archive items besides e-mail messages and attachments, such as Calendar, Tasks, Notes, Persons, Meetings, etc.

Source: Ibis Consulting, Glossary.

Non-Native Environment

A proprietary process in which electronic data is obtained directly from backup tapes without the need to recreate a native environment.

Source: RenewData, Glossary (10/5/2005).

Non-Negative Matrix Factorization

A mathematical technique that summarizes the correlation between items. One of the techniques used in eDiscovery as the basis for concept search, where the items are words.

Source: Herb Roitblat, Search 2020: The Glossary.

Non-Printable Files

Files that can't be printed, such as DLL, EXE, AVI files.

Source: Ibis Consulting, Glossary.

Non-Relevant / Not Relevant

In Information Retrieval, a Document is considered Non-Relevant (or Not Relevant) if it does not meet the Information Need of the search or review effort. The synonym "irrelevant" is rarely used in Information Retrieval.

Source; Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Non-Response Bias

Non-Response Bias occurs when a portion of potential samples is not available for sampling. As an example, if an e-discovery effort is identifying potential responsive engineering documents, and if the documents are in a document format and/or programming language that could not be sampled or understood, there could be a significant non-response Bias. See also, Response Bias.

Source: EDRM Search Glossary.

See also:

Non-Response Bias Response Bias

Normal Distribution

The "bell curve" of classical statistics. The number of Relevant Documents in a Sample tends to obey a Normal (Gaussian) Distribution, provided the Sample size is large enough to capture a substantial number of Relevant and Non-Relevant Documents. In this situation, Gaussian Estimation is reasonably accurate. If the Sample size is insufficiently large to capture a substantial number of both Relevant and Non-Relevant Documents (as a rule of thumb, at least 12 of each), the Binomial Distribution better characterizes the number of Relevant Documents in the Sample, and Binomial Estimation is more appropriate. Also referred to as a Gaussian Distribution.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

In probability theory, the normal (or Gaussian) distribution is a continuous probability distribution. It has a bell-shaped probability density function, known as the Gaussian function or informally as the bell curve, the height of the curve shows the relative likelihood of various values. The area under the curve sums to 1.0, so sections of the curve represent probabilities. The normal distribution derives from the central limit theorem, which says that the average of a large number of random variables is distributed as the normal distribution, however the variables were originally distributed. The normal distribution has wide application in statistics, for example, in sampling.

A graph of the normal distribution. The confidence interval is in the middle in white. The "tails" are shown in yellow. The 95% confidence interval represents 95% of the area under the curve. In a two-tailed distribution, this 95% area is symmetrically aligned around the average of the distribution. Image from https://en.wikipedia.org/wiki/One-_and_two-tailed_tests

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

Gaussian Distribution

NOS

See: Network Operating System (NOS)

NoSQL

Generally interpreted to mean “Not Only SQL,” refers to data bases that are built using structures other than tables and relations. NoSQL databases are typically distributed over many physical machines, horizontally scalable, and are often distributed as open-source software.

Source: Herb Roitblat, Search 2020: The Glossary.

Note Field

A data field that allows the entry of text in a manner similar to word processing software, which is not limited to a specific number of characters. Typically used for attorneys’ notes or comments. A note field cannot be sorted.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Other number field
Production source
Recipient
Subject category
Summary
Text

Notebook Computer

A small laptop computer, usually weighing less than 8 pounds.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Computer
File server
Laptop computer
Microcomputer
Minicomputer
Personal computer
Workstation

Notes Storage Facility (NSF)

Databases in IBM Notes, formerly Lotus Notes, are Notes Storage Facility (.nsf) files, containing basic units of storage known as a "note".

Source: http://en.wikipedia.org/wiki/IBM_Notes.

A Lotus Notes mail container.

Source: Ibis Consulting, Glossary.

A Lotus Notes / Domino database, including email collections.

See also:

Container
EML
Mail container
Mailbox
MSG
Multi-mail container
OST
PST
RFC compliant email
RFC822
Single-mail archive
Single-mail container
SMTP

NSF

See: Notes Storage Facility (NSF)

NT

Refers to Microsoft Windows NT server and workstation software.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

NT Filing System (NTFS)

NTFS (NT file system; sometimes New Technology File System) is the file system that the Windows NT operating system uses for storing and retrieving files on a hard disk. NTFS is the Windows NT equivalent of the Windows 95 file allocation table (FAT) and the OS/2 High Performance File System (HPFS). However, NTFS offers a number of improvements over FAT and HPFS in terms of performance, extendibility, and security.

Source: TechTarget, NTFS (NT file system; sometimes New Technology File System) definition, http://searchwindowsserver.techtarget.com/definition/NTFS

See also:

FAT
File system
NT filing system

NTFS

See: T Filing System (NTFS)

Null Set

The set of Documents that are not returned by a search process or that are identified as Not Relevant by a review process.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Numeric Range Search

A numeric range search is a search for any numbers that fall within a range.

Source: dtSearch Support, Numeric Range Searching, https://support.dtsearch.com/webhelp/dtsearch/default.htm#numeric_.htm

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Phonic search
Phrase search

Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

O

Object

A combination of code and data created at runtime that can be treated as a unit. A table, chart, graphic, equation, or other form of information. See embedded object.

Source: Ibis Consulting, Glossary.

See also:

Bibliographic coding
Embedded object
Link object
Link source
Linked object

Object Linking and Embedding (OLE)

A feature in Microsoft's Windows which allows each section of a compound document to call up its own editing tools or special display features. This allows for combining diverse elements in compound documents.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Objective Coding

The recording of basic data such as date, author, or document type, from documents into a database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Extracting information from electronic documents such as date created, author recipient, CC and linking each image to the information in pre-defined objective fields. In direct opposition to Subjective coding where legal interpretations of data in a document are linked to individual documents. Also called bibliographic coding.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Extracting various segments of information from a document such as its author, recipient, mailing date, or other fields, etc. Objective Coding is usually done from the document text or image because metadata or searchable text may be unavailable (e.g. a handwritten document that has been scanned), or the document may contain inaccurate metadata (e.g. metadata

associated with a document written and signed by a partner might reflect the administrative assistant as the author where the document was originally typed on the assistant's computer).

See also:

Bibliographic Coding	Issue coding	Taxonomic coding
Coding	Level coding	Verbatim coding
Indexing	Subjective coding	
Issue Code	Tag	

Obstruction of Justice

According to Black's law dictionary, obstruction of justice means "impeding or obstructing those who seek justice in a court, or those who have duties or powers of administering justice therein."

Source: RenewData, Glossary (10/5/2005).

Occurrence Count

Occurrence count search allows a legal professional to specify Occurrence count search allows a legal professional to specify that a word appear a certain number of times for the document to be selected.

Source: EDRM Search Glossary.

OCR (Optical Character Recognition)

Optical character recognition is the conversion of a scanned document into searchable text and the rendering of its text susceptible to copying for pasting into a new file. Following the scanning of a given document, OCR software evaluates the scanned data for shapes it recognizes as letters or numerals. OCR technology relies upon the quality of the printed copy and the conversion accuracy of the software. Generally acknowledged to be only 80-85 percent accurate.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing RenewData's Electronic Evidence Reference Chart, http://www.renewdata.com/wall-chart-signup.html

A method of translating printed text and images into a form that a computer can manipulate (into ASCII codes, for example). An OCR system enables you to scan a printed document directly into a computer file.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

A method of scanning printed material and converting it into an electronic file, such as a word-processing file, which can then be searched for specific words or phrases. OCR is distinguishable from "imaging" in that it recognizes only alphanumeric characters and not handwritten or other graphic material.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Software that, in conjunction with a scanner, is able to "recognize" written text and convert it to an ASCII file or import it into a word processor so may perform one of the full text searches.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The computer conversion of scanned input images (bar codes or patterns of bits) to computer recognizable codes (ASCII letters, numbers and characters).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Optical character recognition is a technology which takes data from a paper document and turns it editable text data. The document is first scanned. Then OCR software searches the document for letters, numbers, and other characters.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

When a paper document is scanned into a computer, an image is created. The computer does not recognize the characters of the document as text until OCR software converts the image into text. OCR systems vary widely in the accuracy of their conversion. Even seemingly high accuracy rates can, however, still result in significant numbers of words being misrepresented. A 99% accuracy, for example, would still result in one word out of 20 being misspelled.

See also:

Dirty OCR ICR Pattern recognition

ODBC (Open Database Connectivity)

An application interface from Microsoft that provides a common language between applications and databases on a network.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A database programming interface that provides a common language for Windows applications to access databases on a network.

OEM (Original Equipment Manufacturer)

Classically, a company who buys products from another company, re-labels the products under its own name and re-sells (usually in large quantities). Has come to define nearly any large customer who re-sells products, branded or not.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Offline

When computers and other devices are not connected to the network.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Archival documents stored on optical disks or compact disks that are not connected or installed in the computer, but instead require human intervention to be accessed.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Not connected (to a network).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Offline Storage

The storage of electronic data outside the network in daily use (i.e., on backup tapes) that is only accessible through the off-line storage system, not the network.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

OLE

See: Object Linking and Embedding (OLE)

On-Site Extraction

The extraction of high volumes of data from backup tapes at a client site.

Source: RenewData, Glossary (10/5/2005).

One-Tailed Test

In hypothesis testing, we can be interested in a deviation in either direction or in only one direction. If we are interested in either direction (one score is different from another), we use a two-tailed test. If we are interested in only one direction (one score is less than another), and we don't care if it is greater, then we use a one-tailed test. For example, if we want to know whether a predictive coding system has performed better than chance, then we can use a one-tailed test. We don't care if the predictive coding system is worse than chance (that would not be particularly useful), only if it is better. Confidence intervals can be one-sided or two-sided as well. The tail refers to the yellow regions in the figure.

A graph of the normal distribution. The confidence interval is in the middle in white. The "tails" are shown in yellow. The 95% confidence interval represents 95% of the area under the curve. In a two-tailed distribution, this 95% area is symmetrically aligned around the average of the distribution. Image from https://en.wikipedia.org/wiki/One-_and_two-tailed_tests

Source: Herb Roitblat, Predictive Coding Glossary.

Online

When computers and other devices are connected to the network.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The condition of a computer being connected to a computerized information system such as Lexis. Often refers to being connected to the Internet.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Documents stored on the hard drive or magnetic disk of a computer that are available immediately.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Connected (to a network).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Online Storage

The storage of electronic data as fully accessible information in daily use on the network or elsewhere.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Online Summary

A digest or summary of a document created directly from the computer screen by reading the document and using the cut and paste function to move excerpts to a separate file.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Ontology

A representation of the relationships among words and their meanings that is richer than a Taxonomy. For example, an Ontology can represent the fact that a wheel is a part of a bicycle, that gold is yellow, and so on.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A categorical or conceptual structure that may not be strictly hierarchical (cf. taxonomy). Concepts can be related to one another in complex ways. For example, an ontology may represent that lawyers, paralegals, and judges are associated with one another (one is not strictly a subset of the other).

Source: Herb Roitblat, Search 2020: The Glossary.

Open Database Connectivity

See: ODBC (Open Database Connectivity)

Operating System (OS)

Software which directs the overall activity of a computer (e.g. MS-DOS, Windows, Linux, etcetera).

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

The most important program that runs on a computer. Every general-purpose computer must have an operating system to run other programs. Operating systems perform basic tasks, such as recognizing input from the keyboard, sending output to the display screen, keeping track of files and directories on the disk, and controlling peripheral devices such as disk drives and printers. For large systems, the operating system has even greater responsibilities and powers. It is like a traffic cop -- it makes sure that different programs and users running at the same time do not interfere with each other. The operating system is also responsible for security, ensuring that unauthorized users do not access the system.

Source: Ibis Consulting, Glossary.

Software that controls the operation of a computer.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The software that the rest of the software depends on to make the computer functional. On most PCs this is Windows or the Macintosh OS. Unix and Linux are other operating systems often found in scientific and technical environments.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

DOS	Microsoft Windows	UNIX
Linux	Network operating system	Windows
Microsoft DOS	NOS	Xenix

Optical Character Recognition

See: OCR (Optical Character Recognition)

Optical Disk

Computer media similar to a compact disc that cannot be rewritten. An optical drive uses a laser to read the stored data.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

CD
CD-R
CD-ROM
CD-RW
Disc
Disk
Diskette
DVD
DVD-ROM
Floppy disk
Hard disk
Hard drive
Jaz disk
Laser disc
Magnetic disk
Magnetic storage media
Media
Storage media
WORM disk
Zip disk

Original Equipment Manufacturer

See: OEM (Original Equipment Manufacturer)

OS

See: Operating System (OS)

OST

An offline storage mail container that requires conversion to PST before extraction, in order to be processed as mail messages and their attachments.

Source: Ibis Consulting, Glossary.

Microsoft Outlook Offline file used to save emails.

See also:

Container
EML
Mail container
Mailbox
MSG
Multi-mail container
NSF
PST
RFC compliant email
RFC822
Single-mail archive
Single-mail container
SMTP

Other Number Field

A data field in a database used to capture numbers other than the primary Bates stamp number that appears on the document.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Production source
Recipient
Subject category
Summary
Text

Output

The folder or files created to contain data that results from a given process.

Source: Ibis Consulting, Glossary.

P

PackBits

A compression scheme which originated with the Macintosh. Suitable only for black & white.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Packet

A fixed block of data transmission which also contains identity and routing information.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Page

A single image of a “one piece of paper.” One or several pages make up a “document.”

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Pages Per Minute (PPM)

A measurement of the throughput speed of a scanner - how many letter-size pages the scanner can scan in one minute. Beware: ppm can be misleading.

Source: RSI, Glossary.

Pantone Matching System (PMS)

A color standard in printing.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Paper Discovery

Paper discovery refers to the discovery of writings on paper that can be read without the aid of some devices.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Paper Styles and Definitions

1. Acid Free Paper – Won't change color (yellow) for many years.
2. Brightness – The percentage of light the paper reflects. Most white papers reflect 60% to 90%.
3. Coated Papers – "glossy" paper, coated with clay.
4. Cotton "Rag" Paper – Premium paper with 25% to 100% cotton fibers.
5. Laid finish – Paper surface embossed with lines to resemble handmade paper.
6. Ream – 500 sheets.
7. Vellum finish – A less smooth version of real vellum (fine parchment).
8. Wove finish – Very smooth surface. Characteristic of the majority of papers made.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Parallel

Refers to multiple data bits stored or transmitted simultaneously.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Transmission of all the bits (e.g. in a character) at the same time. If the character has eight bits, there are eight wires. Faster and more expensive than serial where the eight bits would be sent, "sideways", one at a time.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Parallel Port

A parallel port is used for printing because it is faster than a serial port.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Parallel Trial

An Experimental Design for comparing two search or review processes using the same Document Collection and Information Need, in which both processes are applied concurrently but independently, and then the results of the two efforts are compared. (*Cf*. Crossover Trial.)

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Parametric Search

Parameterized search allows searching to be based not on keywords but on certain parameters, such as a document's metadata. Parameterized search is also known as fielded search, because it is frequently performed on data stored within the fields of a database table. Examples include Date Range, Metadata, Custodian, restrictions or promotions based on document tags/review calls.

Source: EDRM Search Glossary.

Parent Document

The primary document in a set of related documents, such as a fax cover sheet or a transmittal letter.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Parent-Child Relationship

In any taxonomy, the superior category can be called a parent, and its subcategories can be called children. An email can be considered, for example, to be a parent of any of its attachments. Conversely, an attachment can be considered to be a child of the email to which it is attached.

Password Protection

The use of personal and confidential identification to allow individual users access to a computer system or specific programs.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Path

The route of directories through which a computer searches to find a particular file. The path name is the full file name, including the name of the directory on which the file is stored.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Pattern Matching

The science of designing computer Algorithms to recognize natural phenomena like parts of speech, faces, or spoken words.

Source: The Grossman-Cormack Glossary of Technology Assisted Review (Version 1.02, Nov. 2102).

Pattern Recognition

An electronic application utilizing an algorithm that searches data for like patterns and flags or extracts the pertinent data. For instance, in looking for addresses, alpha characters followed by a comma and a space followed by two capital alpha characters followed by a space followed by five or more digits are usually the city, state and zip code. By programming the application to look for that pattern, the information can be electronically extracted rather than re-keyed by human intervention.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Dirty OCR

ICR

OCR

Optical Character Recognition

PB (Petabyte)

A petabyte is a measure of computer data storage capacity and is one thousand million million (1,000,000,000,000,000) bytes.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Bit

Byte

KB - kilobyte

MB - megabyte

GB - gigabyte

TB - terabyte

EB - exabyte

PC (Personal Computer)

Technically a computer that conforms to the PC standard set by IBM, the PC now refers to any desktop computer other than a terminal on a Unix system.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

PCI (Peripheral Component Interconnect)

A high-speed interconnect local bus used to support multimedia devices. Promoted by Digital among others.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

PCMCIA (Personal Computer Memory Card International Association)

Plug-in cards for computers (usually portables), which extend the storage and/or functionality.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

PCX (Personal Computer eXchange)

The file format used for drawings by Corel Paint and Windows Paint Brush.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

PDA (Personal Digital Assistant)

Any small hand held wireless device that provides computing and data storage abilities. Examples of PDAs include the Palm Pilot and the BlackBerry.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A hand-held microcomputer that functions like an electronic rolodex and often connects to a larger computer for sharing or transferring information.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A small, usually hand-held, computer which "assists" business tasks.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

PDF (Portable Document Format)

A proprietary format of Adobe Corporation, it has become a de facto standard for transmitting documents that the sender does not want to be altered and for transmitting documents to commercial printers and to the Web for online publishing.

Source: RenewData, Glossary (10/5/2005).

A file format developed by Adobe Systems. PDF captures formatting information from a variety of desktop publishing applications, making it possible to send formatted documents and have them appear on the recipient's monitor or printer as they were intended. To view a file in PDF format, you need Adobe Acrobat Reader, a free application distributed by Adobe Systems.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

A file standard for documents that can be processed (generally viewed and printed) by any computer, regardless of the specific application program which created the original.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

An Adobe technology for formatting documents so that they can be viewed and printed using the Adobe Acrobat reader.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

PDF's can be read using Adobe Acrobat Reader (a free program), regardless of the program used to create the original document. A PDF document can contain text, images, or both. Only PDFs containing text can be searched directly. Those containing images only must be OCRed.

See also:

GIF
Graphic Interchange File
Image file format
Joint photographic expert group
JPEG
Multi-page TIFF
PNG
Portable network graphic
Searchable TIFF
Single-page TIFF
TIFF

Peer-to-Peer Network

A network of computers configured to allow certain files and folders to be shared with everyone or with selected users. Peer-to-peer networks are quite common in small offices that do not use a dedicated file server. All client versions of Windows, Mac and Linux can function as nodes in a peer-to-peer network and allow their files to be shared.

Source: PCMag, Definition of: peer-to-peer network, http://www.pcmag.com/encyclopedia/term/49056/peer-to-peer-network

See also:

Client/server network
LAN - local area network
MAN - metropolitan area network
Network

SAN - storage area network

Stand alone computer

WAN - wide area network

Peripheral

Any hardware device that interfaces with a computer, such as a printer, an external modem, or a scanner. Interfacing may take place through the computer's parallel and serial ports or through a specific interface card.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Peripheral Component Interconnect

See: PCI (Peripheral Component Interconnect)

Personal Computer

See: PC (Personal Computer)

Personal Computer eXchange

See: PCX (Personal Computer eXchange)

Personal Computer Memory Card International Association

See: PCMCIA (Personal Computer Memory Card International Association)

Personal Digital Assistant

See: PDA (Personal Digital Assistant)

Personal Information Manager (PIM)

Software that performs the functions of Rolodex.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Personal Storage File (PST)

There are two types of Outlook Data Files used by Outlook. An Outlook Data File (.pst) is used for most accounts.... Outlook Data Files (.pst) are used for POP3, IMAP, and web-based mail accounts. When you want to create archives or back up your Outlook folders and items on your computer, such as Exchange accounts, you must create and use additional .pst files.... A Personal Folders file (.pst) is an Outlook data file that stores your messages and other items on your computer. This is the most common file in which information in Outlook is saved by home users or in small organizations....

Source: Introduction to Outlook Data Files (.pst and .ost), http://office.microsoft.com/en-us/outlook-help/introduction-to-outlook-data-files-pst-and-ost-HA010354876.aspx.

In Microsoft Outlook, the Personal Folders file (.pst) is a data file that stores all of a user's messages and other items on his/her computer. An Outlook user can create one or more .pst's to organize and back up items for safekeeping. Even when an e-mail system is being run on a Microsoft Exchange Server, Outlook data can be backed up to a .pst file stored either locally on a hard drive or on a network drive -- rather than on the e-mail server. Each .pst file contains all of one's Outlook folders, including the Inbox, Calendar, and Contacts.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Microsoft Office Online, http://office.microsoft.com/en-us/assistance/HA010875321033.aspx

The place where Outlook stores its data (when Outlook is used without Microsoft® Exchange Server). A PST file is created when a mail account is set up. Additional PST files can be created for backing up and archiving Outlook folders, messages, forms and files. The file extension given to PST files is .pst.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A MS Outlook mail container that requires extraction in order to be processed as mail messages and their attachments.

Source: Ibis Consulting, Glossary.

A file used in Outlook to save a collection of emails.

See also:

Container	Multi-mail container	Single-mail archive
EML	NSF	Single-mail container
Mail container	OST	SMTP
Mailbox	RFC compliant email	
MSG	RFC822	

Petabyte

See: PB (Petabyte)

Phase Change

A method of storing information on rewritable optical disks.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Phases (Stages)

Distinct segments of the e-discovery process which contain measurable activities that can be tracked according to volume, cost and time. These segments correspond to the L600 codes for: Identification (L600), Preservation (L610), Collection (L620), Processing (L630), Review (L650), Analysis (L660), Production (L670), Presentation (L680) and Project Management (L690).

Source: EDRM Metrics Glossary

Phonic Search

Phonic searching looks for a word that sounds like the word you are searching for and begins with the same letter.

Source: dtSearch Support, Phonic Searching, https://support.dtsearch.com/webhelp/dtsearch/phonic_s.htm

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Phrase Search

A search consisting of multiple keywords separated by spaces to form a single phrase. For a document to match this search, the entire phrase as entered must be contained within the document.

Source: EDRM Search Guide Glossary.

The search phrase “Massachusetts Mutual” would locate text where the words are side by side.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search

Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Physical Target

When the forensic imaging process targets the entire physical drive or data storage media.

Source: EDRM Collection Standards

Physical Unitization

The assembly of individually scanned pages into documents:

- **Physical unitization** utilizes actual objects such as staples, paper clips and folders to determine pages that belong together as documents for archival and retrieval purposes.
- **Logical unitization** is the process of human review of each individual page in an image collection using logical cues to determine pages that belong together as documents. Such cues can be consecutive page numbering, report titles, similar headers and footers and other logical cues.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Pica

One sixth (1/6) of an inch. Used to measure graphics/fonts.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

PICT (Picture Format)

A color file format exclusively for Macintosh.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Picture Element (Pixel)

Primary unit of color on a computer monitor or in an electronic image.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The basic building block of all images -- a simple dot. In bitonal images, it is merely a black or white dot (see "Bitonal" definition above). In grey scale images, dots will have between 1-to-256 possible values of grey (for an 8-bit grey scale image).

Source: RSI, Glossary.

A dot. One step/addressable position in the total TV or CRT presentation. The minimum VGA display has 307,200 pixels (640 by 480).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A picture element. A pixel is the smallest dot on the screen of a computer display. Screen resolution is usually measured in the number of pixels horizontally and vertically that the screen can display (ranging from 640 × 480 to 1280 × 1024, or higher). Generally, the higher the resolution, that is, the more pixels in the image, the clearer the image will appear.

Picture Format

See: PICT (Picture Format)

Piece of Media (POM)

One unit of physical media (tapes, ZIP/Jaz disks, DLT, HDD, floppy disks, FTP'ed material or e-mailed bundles of files, etc.).

Source: Ibis Consulting, Glossary.

PIM

See: Personal Information Manager (PIM)

Pitch

Characters (or dots) per inch, measured horizontally.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Pivot Table

A PivotTable report is an interactive table, such as that found in Excel that can be used to summarize data. For example, in an expense report, you can use it to sum all of the money spent for meals, or you can summarize how much was spent each day. These tables are interactive, one can rotate its rows and columns to see different summaries of the source data, filter the data by displaying different pages, or display the details for areas of interest.

Pixel

See: Picture Element (Pixel)

Plaintext

The least formatted and therefore most portable form of text for computerized documents.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Platform

The underlying hardware or software for a system. For example, the platform might be an Intel 80486 processor running DOS Version 6.0. The platform could also be UNIX machines on an Ethernet network. The platform defines a standard around which a system can be developed. Once the platform has been defined, software developers can produce appropriate software - and managers can purchase appropriate hardware and applications.

Source: Ibis Consulting, Glossary.

Platform = operating system (or family of operating systems) on the software side. The term cross-platform refers to applications, formats, or devices that work on different platforms. For example, a cross-platform programming environment enables a programmer to develop programs for many platforms at once.

Plug-in

A program that enables a Web browser to present non-HTML documents, such as Adobe Acrobat documents or sound and video programs.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

PMS

See: Pantone Matching System (PMS)

PNG (Portable Network Graphic)

Portable Network Graphics (PNG /ˈpɪŋ/) is a raster graphics file format that supports lossless data compression. PNG was created as an improved, non-patented replacement for Graphics Interchange Format (GIF), and is the most used lossless image compression format on the Internet.

Source: Wikipedia, Portable Network Graphics, https://en.wikipedia.org/wiki/Portable_Network_Graphics

See also:

GIF

Graphic Interchange File

Image file format

Joint photographic expert group

JPEG

Multi-page TIFF

PDF

Portable Document Format

Searchable TIFF

Single-page TIFF TIFF

POD (Print On Demand)

Document images are stored in electronic format and are available to be quickly printed and in the exact quantity required, long or short runs.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Point Estimate

The most likely value for a Population characteristic. When combined with a Margin of Error (or Confidence Interval) and a Confidence Level, it reflects a Statistical Estimate.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Point to Point Protocol (PPP)

A standard for connecting two computers for transferring data.

Pointer

A pointer is an index entry in the directory of a disk (or other storage medium) that identifies the space on the disc in which an electronic document or piece of electronic data resides, thereby preventing that space from being overwritten by other data. In most cases, when an electronic document is “deleted,” the pointer is deleted, which allows the document to be overwritten, but the document is not actually erased.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

An index entry in the directory of a hard disk that identifies the space on the disk where a specific file is located. When a file is “deleted,” it is actually the pointer which is erased and not the file itself.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Policy Integration

A formalized common set of goals and rules that promote cross-functional communication, collaboration, and optimization. Information governance efforts can be crippled by failure to integrate policy.

Source: IGRM White Paper

Polysemy

A single word or expression having multiple meanings.

Source: EDRM Search Glossary.

POM

See: Piece of Media (POM)

Population

The universe of things about which we are trying to infer with our samples. For example, the population may be the set of documents that we want to classify as putatively responsive or putatively non-responsive. The group from which we pull our samples. Also called the sampling frame.

Source: Herb Roitblat, Predictive Coding Glossary.

Port

An interface for connecting peripherals with the computer.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The part of the computer through which a peripheral device may communicate, often a specific type of plug.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Portability

The ability to transport a computer and data from one location to another. Typically a feature of laptop or notebook computers, but also a feature of portable drives or tape systems.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Portable Document Format

See: PDF (Portable Document Format)

Portable Drive

An external disk drive that is plugged into a port on a computer, typically a USB or FireWire port. Typically used for backup, but also as secondary storage. Such units rival internal drives in capacity.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing TechWeb TechEncyclopedia, http://www.techweb.com/encyclopedia/defineterm.jhtml?term=portablehardd rive.

See also:

Disk drive	Magneto-optical drive	Zip drive
Floppy disk drive	Storage device	
Jaz drive	Tape drive	

Portable Network Graphic

See: PNG (Portable Network Graphic)

Portable Volume

A feature that facilitates the moving of large volumes of documents without requiring copying multiple files. Portable volumes enable individual CDs to be easily regrouped, detached and reattached to different databases for a broader information exchange.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Portal

A web site which gives entry to multiple other sites and services.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Portrait Mode

A display where the height exceeds the width.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Portrait Orientation

An image registered so that it is taller than it is wide, with the narrow edge running along top and bottom. When scanning, orientation is determined by the leading edge of the document.

Source: RSI, Glossary.

See also:

Landscape orientation

Positive Agreement

The Probability that, if one reviewer Codes a Document as Relevant, a second independent reviewer will also Code the Document as Relevant. Empirical studies show that Positive Agreement rates of 70% are typical, and Positive Agreement rates of 80% are rare. Positive Agreement should not be confused with Agreement (which is a less informative measure) or Overlap (which is a numerically smaller measure that conveys similar information). Under the assumption that the two reviewers are equally likely to err, Overlap is roughly equal to the

square of Positive Agreement. That is, if Positive Agreement is 70%, Overlap is roughly 70% x 70% = 49%.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Positive Predictive Value (PPV)

See Precision. Positive Predictive Value is a term used in Signal Detection Theory; Precision is the equivalent term in Information Retrieval.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

PPM

See: Pages Per Minute (PPM)

PPP

See: Point to Point Protocol (PPP)

PPV

See: Positive Predictive Value (PPV)

Practice Direction

Practice Directions: these are official adjuncts to the CPR and provide mandated guidance for practitioners in conducting litigation.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

Practice Management System

Also known as Case Management System (CMS). Such systems may include features such as calendar/docket, conflict-checking, document assembly, and maintenance of databases of client and case information.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Precision

Precision measures the number of truly responsive documents in the retrieved set of responsive documents. See also, Recall.

Source: EDRM Search Glossary.

The fraction of Documents identified as Relevant by a search or review effort, that are in fact Relevant. Also referred to as Positive Predictive Value.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Precision is the proportion of retrieved documents that are responsive.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

Recall

Precision-Recall Curve

The curve representing the tradeoff between Precision and Recall for a given search or review effort, depending on the chosen Cutoff value. See Recall-Precision Curve.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Precision-Recall Tradeoff

The notion that most search strategies can be adjusted to increase Precision at the expense of Recall, or vice versa. At one extreme, 100% Recall could be achieved by a search that returned the entire Document Population, but Precision would be low (equal to Prevalence). At the other extreme, 100% Precision could be achieved by a search that returned a single Relevant Document, but Recall would be low (equal to 1/N, where N is the number of Relevant Documents in the Document Population). More generally, a broader search returning many Documents will have higher Recall and lower Precision, while a narrower search returning fewer Documents will have lower Recall and higher Precision. A Precision-Recall Curve illustrates the Precision-Recall Tradeoff for a particular search method.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Predictive Coding

An industry-specific term generally used to describe a Technology-Assisted Review process involving the use of a Machine Learning Algorithm to distinguish Relevant from Non-Relevant Documents, based on a Subject Matter Expert's Coding of a Training Set of Documents. See Supervised Learning and Active Learning.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A group of machine learning technologies that predict which documents are and are not responsive based on the decisions applied by a subject matter expert to a small sample of documents.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

CAR TAR

Presentation Phase

Displaying ESI before audiences (at depositions, hearings, trials, etc.), especially in native and near-native forms, to elicit further information, validate existing facts or positions, or persuade and audience.

Source; EDRM Stages

Corresponds to UTBMS Code L680. Activities and actions to prepare and display ESI before audiences (at depositions, hearings, trials, etc.), especially in native and near-native forms, to elicit further information, validate existing facts or positions, or persuade an audience.

Source: EDRM Metrics Glossary

Preservation Phase

Ensuring that ESI is protected against inappropriate alteration or destruction.

Source: EDRM Stages

Corresponds to UTBMS Codes L610-L619. Preservation Order, Legal Hold, Quality Assurance and Control.

Source: EDRM Metrics Glossary

Prevalence

The fraction of Documents in a Population that are Relevant to an Information Need. Also referred to as Richness or Yield.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The richness or proportion of responsive documents in a collection. More broadly, the prevalence refers to the proportion of one kind of item in a population of items.

Source: Herb Roitblat, Predictive Coding Glossary.

Principal Component Analysis

A mathematical technique that summarizes the correlation between items. One of the techniques used in eDiscovery as the basis for concept search, where the items are words.

Source: Herb Roitblat, Search 2020: The Glossary.

Print On Demand

See: POD (Print On Demand)

Private Network

A network that is connected to the Internet but is isolated from the Internet.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Privilege

A special and exclusive legal advantage or right. Examples include attorney work product and certain communications between an individual and his or her attorney, which are protected from disclosure.

Privilege Log

A record of the responsive and/or relevant documents that are being withheld from production on a claim that they either contain attorney-client communication or are attorney work-product. Though there is not standard rule describing the necessary content for a privilege log, the Federal Rules of Civil Procedure contain a general requirement that a privilege log "describe the nature" of the privileged document in a manner that "will enable other parties to assess the claim." Fed. R. Civ. P. 26(b)(5)(A).

Source: EDRM Metrics Glossary

A list of a set of documents that a Producing Party did not produce on account of Privilege such as Attorney-Client Privilege.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Privileged Documents

A set of documents that a Producing Party is not required to provide, since they fall into Privilege such as Attorney-Client Privilege. The existence of such documents should be recorded in the Privilege Log.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Probabilistic Latent Semantic Analysis

A variant of Latent Semantic Analysis based on conditional Probability rather than on correlation.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A statistical procedure for finding the underlying dimensions of correlated terms. Like Latent Semantic Analysis, this procedure attempts to capture the meaning shared by multiple terms to provide a concept search capability. It differs some from LSA in that it involves a different statistical model. Also called probabilistic latent semantic indexing.

Source: Herb Roitblat, Predictive Coding Glossary.

Probabilistic Model

A class of mathematical models that are described in the language of probability without necessarily involving randomness. For example, if we find that three times out of four, when a certain word is used in a document, the document is responsive, then a probabilistic model will include an estimate that the probability that a document containing that word (all other things being equal) is more probably responsive (75%) than non-responsive (25%).

Source: Herb Roitblat, Search 2020: The Glossary.

Probability

The fraction (The fraction of a set of Documents having some particular property (typically Relevance).) of times that a particular outcome would occur, should the same action be repeated under the same conditions an infinite number of times. For example, if one were to flip a fair coin, the Probability of it landing “heads” is one-half, or 50%; as one repeats this action indefinitely, the fraction of times that the coin lands “heads” will become indistinguishable from 50%. If one were to flip two fair coins, the Probability of both landing “heads” is one-quarter, or 25%.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Process Transparency

he shared ownership and execution of Information Governance processes ensuring that accountabilities and dependencies across the stakeholders are clearly defined by each group to promote efficient and effective management of information.

Source: IGRM White Paper

Processing Phase

Reducing the volume of ESI and converting it, if necessary, to forms more suitable for review and analysis.

Source: EDRM Stages

Corresponds to UTBMS Code L630-L639. ESI Stage, Preparation and Process, Scanning - Hard Copy, Foreign Language Translation, Exception Handling, Quality Assurance and Control.

Source: EDRM Metrics Glossary

Producing Party

A party that owns the complete collection of ESI, and is responsible for producing a portion of the ESI that is deemed to be relevant for a legal case or legal enquiry.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Production

Delivering ESI to others in appropriate forms & using appropriate delivery mechanisms.

Source: EDRM Stages.

Delivery of data or information in response to an interrogatory, subpoena or discovery order or a similar legal process.

Source: RenewData, Glossary (10/5/2005).

Production De-Duplication

Culling of a document if multiple copies of that document reside within the same production set. For example, if two identical documents are both marked responsive, non-privileged, production de-duplication ensures that only one of those documents are produced. Contrast with case de-duplication and custodian de-duplication.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Basic de-duplication
Case de-duplication
Custodian de-duplication
De-duplication
Duplicate
Dynamic de-duplication
Global Deduplication
Horizontal Deduplication
Vertical Deduplication

Production Phase

Delivering ESI to others in appropriate forms and using appropriate delivery mechanisms.

Source: EDRM Stage

Corresponds to UTBMS Codes L670-L679. Conversion of ESI to Production Format, Quality Assurance and Control.

Source: EDRM Metrics Glossary

Production Source

A data field in a database that records the individual or company that produced the particular document.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Recipient
Subject category
Summary
Text

Program

A series of instructions to the computer.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The term for a software application.

Source; Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Project Management Phase

Corresponds to UTBMS Code L690. Activities or actions to associated with supervising or managing specific activities or actions throughout the EDRM continuum such as conducting meetings and team calls, developing work plans, budgets, forecasts, reports and other

meaningful activities or for general project management not associated with a particular "L" code.

Source: EDRM Metrics Glossary

Project Manager

An individual responsible for administration and supervision over a particular database or automation project.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Prompt

A display that asks the operator to perform a specific action.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The DOS prompt.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Proportion

The fraction of a set of Documents having some particular property (typically Relevance).

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Proportionality

Pursuant to Federal Rules of Civil Procedure 26(b)(2)(B), 26(b)(2)(C), 26(g)(1)(B)(iii), and other federal and state procedural rules, the legal doctrine that Electronically Stored Information may be withheld from production if the cost and burden of producing it exceeds its potential value to the resolution of the matter. Proportionality has been interpreted in the case law to apply to preservation as well as production.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The overriding objective of the CPR is to enable the court to deal with cases justly (CPR1). Specifically this is stated to include "dealing with the case in ways which are proportionate..." and it then goes on to list factors which need to be considered to determine what is proportionate. Taken together these factors are generally referred to together as Proportionality.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

Proximity Search

A Proximity Search searches for multiple keywords. The matching documents must contain all the keywords, with the keywords occurring within a specified number of words from each other.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Retrieves a word only when it occurs within a specific number of lines or words of another word.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

For "full-text" searches, the ability to look for words which are within a prescribed distance of another word (e.g. find "glove" within 15 words of "baseball".)

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

PST

See: Personal Storage File (PST)

Q

QA & Control

A common element within each e-discovery Phase which refers to defined steps, procedures and methods taken to ensure that work is done completely, accurately and in a manner which is consistent with expectations, instructions and best practices.

Source: EDRM Metrics Glossary

QBIC

See: Query By Image Content (QBIC)

QIC

See: Quarter Inch Cartridge (QIC)

Quad-Gram

An N-Gram where N = 4 (i.e., a 4-gram).

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Quality Assurance

A method to ensure, after the fact, that a search or review effort has achieved reasonable results.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Quality Control

Ongoing methods to ensure, during a search or review effort, that reasonable results are being achieved.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Any process used to check the accuracy and consistency of information coded into a database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The process of ensuring the highest level of results in a given task. In document management processes, this includes image quality (resolution, skew, speckle, legibility); data quality (correct information in appropriate fields, validated data for dates, addresses, names/issues lists).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Quality Control / Quality Assurance

Process of validation during post selection of data; throughout review, pre-production to identify inconsistencies in document productions, to test for conflicting review calls.

Source: EDRM Search Glossary.

Quarter Inch Cartridge (QIC)

Digital recording tape, 2000 feet long, with an uncompressed capacity of 5 GB.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Backup	Digital audio tape	Media
Backup tape	Disaster recovery tape	Tape
DAT - digital audio tape	DLT - digital linear tape	
Data extraction	Magnetic storage media	

Query

A formal search command provided as input to a search tool.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Ask for information or data.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A request for data sent to a database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A search request in a database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A query is a request to a search engine or similar information retrieval system. Queries may consist of key words, phrases, complex expressions, or even whole documents.

Source: Herb Roitblat, Search 2020: The Glossary.

Query By Image Content (QBIC)

An IBM search system for stored images which allows the user to sketch an image and then search the images files to find those which most closely match. The user can specify color and texture – such as sandy beaches or clouds.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Query Expansion

The process of adding Search Terms to a Query to improve Recall, often at the expense of decreased Precision.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A process wherein a query submitted by a user is modified to include additional terms. The expanded query may include synonyms of the initial query, spelling alternatives, or other related words. Query expansion is one of the methods used to support concept search.

Source: Herb Roitblat, Search 2020: The Glossary.

R

R-Squared (R^2)

A statistical measure indicating how good one term is at predicting another. Perfect predictions would result in R^2 values of 1.0. If one term is useless for predicting the other, then R^2 would be 0.0. The higher the R^2, the better the prediction. More technically, R^2 measures how well variability in one term predicts the variability in the other.

Source: Herb Roitblat, Predictive Coding Glossary.

R^2

See: R-Squared (R^2)

RAID (Redundant Array of Independent Disks)

Arrays or Jukeboxes of CD-ROM's or CD-R's. There are five commonly used, different levels of data protection, RAID 1 through RAID 5, which are tradeoffs of protection versus storage capacity. These include:

- Level 0: Data written in blocks across multiple drives without an protection on failures.
- Level 1: Disk Mirroring.
- Level 3: The drive spindles are synchronized such that the heads all seek at the same time and are positioned over the same read/write areas simultaneously. Data is written one bit at a time with parity to a separate drive. Thus if there were four disks in the array and there was a megabyte of data to transferred at 1 MB/sec, the effective rate is 4MB/sec.
- Level 5: Writes data in chunks (usually smaller blocks 512 bytes to 2 K) with the parity striped along with the data. Achieves a higher I/O rate.

Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

RAM (Random Access Memory)

The hardware inside a computer that retains memory on a short-term basis and stores information while the user utilizes the computer.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

The main memory of the computer, where active software and temporary files are stored and most of the computer's work is performed. Data stored in RAM are temporarily stored and are lost when the computer is turned off.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Memory which can be read or written in any section with one instruction sequence.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

The working memory of the computer into which application programs can be loaded and executed.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

DRAM Memory ROM

RAND Study

A 2012 study (Nicholas M. Pace & Laura Zakaras, *Where the Money Goes: Understanding Litigant Expenditures for Producing Electronic Discovery*, RAND Institute for Civil Justice (2012)), indicating that Document review accounts for 73% of Electronic Discovery costs, and concluding that "[t]he exponential growth in digital information, which shows no signs of slowing, makes a computer-categorized review strategy, such as predictive coding, not only a cost-effective choice but perhaps the only reasonable way to handle many large-scale productions."

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Random

Unpredictable. Random selection means that each item has an equal chance of being selected and there is no systematic bias to select one item rather than another. Coin flips are random. Knowing that one coin flip came up heads does not change the likelihood that the next coin flip will come up heads (these coin flips are said to be independent).

Source: Herb Roitblat, Predictive Coding Glossary.

Random Access Memory

See: RAM (Random Access Memory)

Random Sample / Random Sampling

A subset of the Document Population selected by a method that is equally likely to select any Document from the Document Population for inclusion in the Sample; the Sample resulting from such action. Random Sampling is the basis of Statistical Estimation.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The statistical process of choosing objects randomly, meaning that each object has an equal chance of being selected. Random sampling can be used to train predictive coding systems and to evaluate their efficacy.

Source: Herb Roitblat, Predictive Coding Glossary.

Range Search

A database query within a certain range of dates or document numbers.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Raster

Represents images by a horizontal and vertical array of dots or pixels. A method of representing an image with a grid (or “map”) of dots or pixels. Typical raster file formats are GIF, JPEG, TIFF, PCX, BMP, etc.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

RAW Image File

A RAW image file is a bit-by-bit copy of data on a disk or volume, without additions, deletions, or metadata. Originally used by dd, the RAW image format is supported by most computer forensic applications.

Source: http://www.forensicswiki.org/wiki/Raw_Image_Format

RDBMS (Relational Database Management System)

Relational Database Management System. This is a technical term for the class of software programs that manage data using a relational schema, such as Microsoft SQL Server or Oracle.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Read Only Memory (ROM)

Read-only memory (ROM) is a type of non-volatile memory used in computers and other electronic devices. Data stored in ROM can only be modified slowly, with difficulty, or not at all, so it is mainly used to store firmware (software that is closely tied to specific hardware and unlikely to need frequent updates) or application software in plug-in cartridges.

Source: Wikipedia, Read-only memory, https://en.wikipedia.org/wiki/Read-only_memory

See also:

DRAM Memory RAM

Reboot

To stop and start the operating system again. Usually done when a problem occurs or the computer “locks up” and is accomplished by pressing the Control, Alternate, and Delete keys at the same time.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Recall

Recall measures the number of responsive documents retrieved compared to the total number of responsive documents in the corpus. Recall cannot be absolute unless all documents have been searched and all have been reviewed. Since Recall measures the ratio of responsive documents against the full corpus, the number of responsive documents in the corpus is difficult to determine. See the EDRM Search Guide regarding precision, recall, and sampling for more information. See also, Precision.

Source: EDRM Search Glossary.

The fraction of Relevant Documents that are identified as Relevant by a search or review effort.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The proportion of responsive documents in the entire collection that have been retrieved.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

Precision

Recall-Precision Curve

The curve representing the tradeoff between Recall and Precision for a given search or review effort, depending on the chosen Cutoff value.

Source: The Grossman-Cormack Glossary of Technology Assisted Review (Version 1.02, Nov. 2102).

Recall-Precision Graph

A graph that shows the tradeoff between precision and recall. Typically, the higher the recall level, the lower the precision level. In order to get more of the responsive documents, one usually has to accept more irrelevant documents.

Source: Herb Roitblat, Predictive Coding Glossary.

Receiver Operating Characteristic Curve (ROC)

In Signal Detection Theory, a graph of the tradeoff between True Positive Rate and False Positive Rate, as the Cutoff is varied.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Recipient

A data field containing the name of the individual or company who received a specific document. Also called "addressee" field.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field

Attorney notes field

Author field

Beginning document number

Beginning number field	End document number	Other number field
Copyee field	Field	Production source
Cross-reference field	Index/coding field	Subject category
Customized data field	Key field	Summary
Customized field definition	Marginalia	Text
Data field definition	Names mentioned in text	
Date field	Note field	

Record

Information, regardless of medium or format, that has value to an organization. Collectively the term is used to describe both documents and electronically stored information.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A collection of related fields or items of data, treated as a unit. For example, each listing in a Personal Information Manager is a record.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

An individual item in a document database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Record Level Deletion

Deletion is the process whereby data is removed from active files and other data storage structures on computers and rendered inaccessible except using special data recovery tools designed to recover deleted data. Deletion occurs in several levels on modern computer systems:

1. File level deletion: Deletion on the file level renders the file inaccessible to the operating system and normal application programs and marks the space occupied by the file's directory entry and contents as free space, available to reuse for data storage.
2. Record level deletion: Deletion on the record level occurs when a data structure, like a database table, contains multiple records; deletion at this level renders the record inaccessible to the database management system (DBMS) and usually marks the space occupied by the record as available for reuse by the DBMS, although in some cases the space is never reused until the database is compacted. Record level deletion is also characteristic of many e-mail systems.
3. Byte level deletion: Deletion at the byte level occurs when text or other information is deleted from the file content (such as the deletion of text from a word processing file); such deletion may render the deleted data inaccessible to the application intended to be used in processing the file, but may not actually remove the data from the file's

content until a process such as compaction or rewriting of the file causes the deleted data to be overwritten.

Source: Merrill Corporation, Electronic Discovery Glossary.

Deletion is the process whereby data is removed from active files and other data storage structures on computers and rendered inaccessible except using special data recovery tools designed to recover deleted data.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Removing active files making them unavailable. Special data recovery tools can still retrieve these files.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Record Lifecycle

The time period from when a record is created until it is disposed.

Source; Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Records Custodian

A records custodian is an individual responsible for the physical storage and protection of records throughout their retention period. In the context of electronic records, custodianship may not be a direct part of the records management function in all organizations.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Records Management

Records Management is the planning, controlling, directing, organizing, training, promoting and other managerial activities involving the lifecycle of information, including creation.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The process of maintaining business documents or records. A records management plan includes policies for document retention and destruction. Records management plans are often designed by a collaboration among information technology, business units, and legal departments.

Records Retention Period

The length of time a given records series must be kept, expressed as either a time period (i.e., four years), an event or action (i.e., audit), or a combination (i.e., six months after audit).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Records Retention Schedule

A plan for the management of records, listing types of records and how long they should be kept; the purpose is to provide continuing authority to dispose of or transfer records to historical archives.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Red, Green and Blue (RGB)

The three primary colors in the additive color family which create all the computer color video signals for a computer's color terminal.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Redact

A portion of the image is blacked out intentionally to conceal information from the document.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

The process of removing privileged information from a document. This is usually accomplished by placing a black area over the privileged text.

Reduced Instruction Set Chip (RISC)

A type of computer chip that combines many instructions in order to speed up processing.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Redundant Array of Independent Disks (RAID)

See: RAID (Redundant Array of Independent Disks)

Refresh Rate

How many times a second and image on a CRT or TV is updated.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Region

An area of an image file that is selected for specialized processing. Also called a “zone.”

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Registration

Lining up a forms image to determine which fields are where. Also, entering pages into a scanner such that they are correctly read.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Registry

The system configuration files used by Microsoft Windows to store settings about user preferences, installed software, hardware and drivers and other settings required for Windows to run correctly.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Regular Expressions

A pattern that describes what the search should return based on special characters added to the keyword. For example, car* uses the character * as a wildcard, and the resulting documents should contain words that begin with the characters “car”, such as car, cartoon, or cartography.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Related Word Search

Related words search allows a legal professional to specify a word and other words that are deemed to be related to it. Typically, such related words are determined as either part of concept search or by statistical co-occurrence with other words.

Source: EDRM Search Glossary.

Relational Database

A relational database is a collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables. Invented by E. F. Codd at IBM in 1970.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing SearchDatabase.com, http://searchoracle.techtarget.com/sDefinition/0,,sid41_gci212885,00.html.

A database in which some items in one type of record refer to items in another type of record. Relational databases generally link together two or more tables or files from different databases through a common field or within ranges, thus allowing searches of multiple fields, such as dates.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A database containing records in fields that are somehow connected or "related." This allows simultaneous searches of multiple fields.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A style of data storage and access where the data are stored in tables. Each row contains one record, and each column contains one variable for that record. Relational databases also allow references (relations) between tables. SQL, structured query language is the typical method used to access the information in relational databases.

Source: Herb Roitblat, Search 2020: The Glossary.

See also:

Database
Flat file database
Full text database
SQL
WAIS - wide area information server

Relational Database Management System

See: RDBMS (Relational Database Management System)

Relevance / Relevant

In Information Retrieval, a Document is considered Relevant if it meets the Information Need of the search or review effort.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Relevance Feedback

An Active Learning process in which the Documents with the highest likelihood of Relevance are coded by a human, and added to the Training Set.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A class of machine learning techniques where users indicate the relevance of items that have been retrieved for them and the machine learns thereby to improve the quality of its recommendations.

Source: Herb Roitblat, Search 2020: The Glossary.

Source: Herb Roitblat, Predictive Coding Glossary.

Relevance Ranking

A search method in which the results are ranked from the most likely to the least likely to be Relevant to an Information Need; the result of such ranking. Google Web Search is an example of Relevance Ranking.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Relevancy Rank

A measurement of relevancy of a document, so that the Search Hits within a Search Results can be ordered. Relevancy measurements often involve counting the number of occurrences of a keyword within a document, as well as number of documents a keyword is found in.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Remote Connectivity

The use of a computer outside the user's office. Commonly associated with the use of portable laptop or notebook computers, but may also refer to the ability to access computers from other offices, from the courtroom, or from the client's office.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Render

The process of converting responsive documents into a standard format typically TIFFs or PDFs. These documents are historically delivered on paper, though they may be produced as images or as original electronic files depending on the case, requests of the attorneys, etc.

Report

The use of a computer outside the user's office. Commonly associated with the use of portable laptop or notebook computers, but may also refer to the ability to access computers from other offices, from the courtroom, or from the client's office.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Repository

A centralized database stored on a computer that houses specific information.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Repository for Electronic Records is a direct access device on which the electronic records and associated metadata are stored. Sometimes called a “records store,” “online repository” or “records archive.”

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Request for Admission

In a civil action, a request for admission is a discovery device that allows one party to request that another party admit or deny the truth of a statement under oath. If admitted, the statement is considered to be true for all purposes of the current trial. Parties may also use this discovery device to request that other parties verify that documents are genuine.

Source: Legal Information Institute Wex, Requests for admission, https://www.law.cornell.edu/wex/requests_for_admission

External links:

Rule 36. Requests for Admission, https://www.law.cornell.edu/rules/frcp/rule_36

See also:

Discovery request Document request Interrogatory

Request for Comments (RFC)

The means by which internet standards are created and modified. RFCs are distributed by technical experts acting on their own initiative and reviewed by the Internet at large, rather than formally promulgated through an institution such as ANSI. For this reason, they remain known as RFCs even once adopted as standards.

Request for Production of Documents

See: Document Request

Requesting Party

A party that does not own the ESI and is requesting that the Producing Party which owns the ESI to provide some subset of the ESI based on a Search Request.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Residual Data

Sometimes referred to as "ambient data," refers to data that is not active on a computer system. Residual data includes:

Data found on media free space;

Data found in file slack space; and

Data within files that has functionally been deleted in that it is not visible using the application with which the file was created, without use of undelete or special data recovery techniques.

Source: Merrill Corporation, Electronic Discovery Glossary.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Data that is not currently live on the computer system, including data found in file slack space, data found on media free space, and data from deleted files. Also known as "ambient data."

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Data that is not active on a computer system such as data in media free space, slack space or files that have been “deleted”. Sometimes called "ambient data."

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ambient data | Free space | Swap file

Fragmented data | Slack space | Unallocated space

Resolution

Indicates the number of dots, often measured in dpi, that make up an image on a screen or printer. The larger the number of dots, and thus the higher the resolution, the finer and smoother images can appear when displayed at a given size. Low resolution causes jagged characters. The ideal resolution is a trade-off between quality and the overhead in storage power and processing strength required to use it.

Source: RSI, Glossary.

The visual clarity of a display screen or printer.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Response Bias

One type of Response Bias can occur if the sampling process considers the content of the documents. See also, Non-Response Bias.

Source: EDRM Search Glossary.

See also:

Non-Response Bias

Responsive File

A file that is responsive to one of the filters (full text search, extension/size, date, MD5-known, cookies, sender/recipient, and custom processing) in an electronic discovery process.

Source: Ibis Consulting, Glossary.

See also:

Responsive/Relevant Documents

Responsive/Relevant Documents

A subset of ESI that matches potentially the desired set of documents for the case.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

See also:

Responsive/Relevant Documents

Responsiveness

A Document that is Relevant to an Information Need expressed by a particular request for production or subpoena in a civil, criminal, or regulatory matter.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A standard that measures whether a document fits the established parameters of the document request.

Restore

A method of preparing a data set for processing by converting mail backups to mail archives (for example, PST for MS Outlook and NSF for Lotus Notes).

Source: Ibis Consulting, Glossary.

In data management, restore is a process that involves copying backup files from secondary storage (tape, Zip disk or other backup media) to hard disk. A restore is performed in order to return data to its original condition if files have become damaged, or to copy or move data to a new location.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

To transfer data from a backup medium (such as tapes) to an on-line system, often for the purpose of recovery from a problem, failure, or disaster. Restoration of archival media is the transfer of data from an archival store to an on-line system for the purposes of processing (such as query, analysis, extraction or disposition of that data). Archival restoration of systems may require not only data restoration but also replication of the original hardware and software operating environment. Restoration of systems is often called “recovery”.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Whatis.com, http://searchstorage.techtarget.com/sDefinition/0,,sid5_gci965124,00.html.

Retention Period

See: Records Retention Period

Retention Schedule

See: Records Retention Schedule

Retrieval

The on-screen result of a query.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Review Feedback Validation

Review feedback validation involves cross referencing the results of search with the calls made by attorneys during document review. The document level classification as relevant or privileged provides keen insight into refining the search and selection criteria or in identifying gaps that require additional analysis. This feedback will be used for additional analysis and to refine the Search Criteria sets. The feedback may identify categories of documents that are not yielding responsive documents and or could identify documents to be excluded from the review set. Also, the feedback may identify new categories of documents that should be included and the criteria will be broadened to include those documents in the review set.

Source: EDRM Search Glossary.

Review Phase

Evaluating ESI for relevance and privilege.

Source: EDRM Stages

Corresponds to UTBMS Code L650-L659. Hosting Costs, Review Planning and Training, Objective and Subjective Coding, First Pass Document Review, Second Pass Document Review, Privilege Review, Redaction, Quality Assurance and Control.

Source: EDRM Metrics Glossary

Rewritable

Storage devices where the data may be written more than once – typically hard drives, floppies and optical disks. The assets are re-use, high speed and capacity. The optical disks have the same basic characteristics as a CD-ROM, except that you can write over the existing data.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

RFC

See: Request for Comments (RFC)

RFC Compliant Email

Emails that are consistent with the internet standards for such documents. The standards are established through a request for comments resulting in a general and open discussion. Email RFCs include RFC 1939 -- Post Office Protocol, and RFC 2821; Simple Mail Transfer Protocol. Compliance to these protocols ensures that the email can be processed accurately.

See also:

Container	Multi-mail container	Single-mail archive
EML	NSF	Single-mail container
Mail container	OST	SMTP
Mailbox	PST	
MSG	RFC822	

RFC822

The standard for the format of ARPA Internet text messages. This is a universal (and outdated) standard for e-mail that is entirely text-based, portable and readable by virtually any system.

Source: Ibis Consulting, Glossary.

See also:

Container	Multi-mail container	Single-mail archive
EML	NSF	Single-mail container
Mail container	OST	SMTP
Mailbox	PST	
MSG	RFC compliant email	

RGB

See: Red, Green and Blue (RGB)

Richness

The proportion or prevalence of responsive documents in a collection.

Source: Herb Roitblat, Predictive Coding Glossary.

See Prevalence or Yield.

Source: The Grossman-Cormack Glossary of Technology Assisted Review (Version 1.02, Nov. 2102).

RISC

See: Reduced Instruction Set Chip (RISC)

ROC

See: Receiver Operating Characteristic Curve (ROC)

Rollback

The functionality to undo application processes.

Source: Ibis Consulting, Glossary.

Rolling Collection / Rolling Ingestion

A process in which the Document Collection is periodically augmented as new, potentially Relevant Documents are identified and gathered. Whenever the Document Collection is augmented, the results of prior search or review efforts must be supplemented to account for the new Documents.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Rolling Production

A process in which Responsive Documents are delivered incrementally to a requesting party to provide timely, partial satisfaction of a Document request.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

ROM

See: Read Only Memory (ROM)

Rotary Camera

In microfilming, the papers are read "on the fly" with a camera that's synchronized to the motion.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Router

A piece of hardware that routes data from a local area network (LAN) to a phone line.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Rule

A formal statement of one or more criteria used to determine a particular outcome, e.g., whether to Code a Document as Relevant or Non-Relevant.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Rule Base

A set of Rules created by an expert to emulate the human decision-making process for the purposes of Classifying Documents in the context of E-Discovery.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Rule-Based Workflow

A programmed series of automated steps that route documents to various users on a multi-user imaging system.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Ad Hoc Workflow Rule-Based Workflow Workflow

S

S-HTTP (Secure HTTP)

Secure HTTP or S-HTTP enables the secure exchange of information and files on the Web. S-HTTP files are encrypted and/or contain a digital certificate. This type of transaction security is likely to be used by financial institutions, because S-HTTP is more secure than a userID and password.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Sample / Sampling

A subset of the Document Population used to assess some characteristic of the Population; the act of generating such a subset of the Document Population. See Interval Sample, Judgmental Sample, Statistical Estimate, Statistical Sample, or Systematic Sample.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The process of selecting a subset of items from a population and inferring from the characteristics of the sample what the characteristics of the population are likely to be. Often refers to a simple random sample, which each item in the population has an equal chance of being selected in the sample.

Source: Herb Roitblat, Predictive Coding Glossary.

Sample Size

The number of documents drawn at random that are used to calculate a Statistical Estimate.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Sampling

Sampling is a method of reviewing statistical ratios of complete or portions of a classified corpus for the purposes of validation.

Source: EDRM Search Glossary.

The process of statistically testing data for the presence of relevant information. Often used to provide courts with a cost estimate in order to allocate cost sharing.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Sampling usually (but not always) refers to the process of statistically testing a database for the likelihood of relevant information. It can be a useful technique in addressing a number of issues relating to litigation, including decisions what repositories of data are appropriate to search in a particular litigation, and determinations of the validity and effectiveness of searches or other data extraction procedures. Sampling can be useful in providing information to the court about the relative cost burden versus benefit of requiring a party to review certain electronic records.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Source: Merrill Corporation, Electronic Discovery Glossary.

Sampling Distribution

The probability distribution of a given measure based on a random sample. Some values are more likely than others. The sampling distribution tells about the likelihood or probability of each value. We can use sampling distributions to test hypotheses without having to compute for ourselves all possible combinations of the events that lead to the outcome. The most commonly used sampling distribution is the so-called normal or Gaussian distribution, the familiar bell-shaped curve. Values near the center of the distribution, the mean or average, are

more likely than values that are far away from the center. When drawn, the sampling height of the sampling distribution shows the probability of that value. The area under the curve tells us about the probability of scores covered by that area.

Source: Herb Roitblat, Predictive Coding Glossary.

Sampling Frame

See: Population

Sampling Rate

The frequency at which analog signals are converted to digital values during digitization. The higher the rate, the more accurate the process. In printing The number of pixels scanned per half tone dot.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

SAN (Storage Area Network)

A storage-area network (SAN) is a dedicated high-speed network (or subnetwork) that interconnects and presents shared pools of storage devices to multiple servers.

Source: TechTarget, storage area network (SAN) definition, http://searchstorage.techtarget.com/definition/storage-area-network-SAN

A storage area network (SAN) is a network which provides access to consolidated, block level data storage. SANs are primarily used to enhance storage devices, such as disk arrays, tape libraries, and optical jukeboxes, accessible to servers so that the devices appear to the operating system as locally attached devices. A SAN typically has its own network of storage devices that are generally not accessible through the local area network (LAN) by other devices. The cost and complexity of SANs dropped in the early 2000s to levels allowing wider adoption across both enterprise and small to medium-sized business environments.

Source: Wikipedia, Storage area network, https://en.wikipedia.org/wiki/Storage_area_network

See also:

Client/server network

LAN - local area network

MAN - metropolitan area network

Network

Peer-to-peer network

Stand alone computer

WAN - wide area network

Sanctions

Consequences, punishments, and penalties imposed by the court for violation of the rules or orders of the court, or by regulators for violations of the rules or orders of regulatory bodies.

Source: Ibis Consulting, Glossary.

Sandbox

A network or series of networks that are not connected to other networks.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Sans Serif

Without serif.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Scalability

The ability of a system to add hardware to increase power or performance without requiring any adjustments to the underlying system.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The capacity of a system to expand without requiring major reconfiguration or re-entry of data. Multiple servers or additional storage can be easily added.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Scale to Gray

An option to display a black and white image file in an enhanced mode, making it easier to view. A scale-to-gray display uses gray shading to fill in gaps or jumps (known as aliasing) that occur when displaying an image file on a computer screen. Also known as grayscale.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Scan

The process of creating an electronic image of a paper document, usually for the purpose of loading into a litigation support system.

Source: LitSavant Ltd., Glossary, http://www.litsavant.com/full-glossary.aspx

Scanning is the process of converting a hard copy paper document into a digital image for use in a computer system. After a document has been scanned, it can be reviewed using field and full-text searching, instant document retrieval, and a complete range of electronic document review options.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Scanner

An input device commonly used to convert paper documents into computer images. Scanner devices are also available to scan microfilm and microfiche.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Double-sided scanner
Duplex scanner
Flatbed scanner
Simplex scanner

Scanning

See: Scan

Scanning Software

Software that enables a scanner to deliver industry standard formats for images in a collection. Enables the use of coding of the images. IPRO, DocuLex and ZyImage are several examples.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

SCSI (Small Computer Systems Interface)

Pronounced "scuzzy". An industry standard (of sorts) for connecting peripheral devices and their controllers to a microprocessor. SCSI defines both hardware and software standards for communication between a host computer and a peripheral. Computers and peripheral devices designed to meet SCSI specifications should work together.

Source: RSI, Glossary.

Pronounced “skuzzy.” A standard for attaching peripherals (notably mass storage devices and scanners) to computers. SCSI allows for up to 7 devices to be attached in a chain via cables. The current SCSI standard is “SCSI II,” also known as “Fast SCSI.”

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

SCSI Scanner Interface

A device used to connect a scanner with a computer.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Search

A method of finding terms within data sets. Search types include Boolean connectors and special characters to define a search. Types include: natural language (simple words or

phrases), Boolean operators, and searches introduced by special characters for wildcard, stemming, fuzzy, phonic, synonym, numeric range and variable weighting searches.

Source: Ibis Consulting, Glossary.

The ability to look within the data and search by a name, date or keyword to find desired information.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A database query.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The process of locating and identifying documents that are relevant. In information retrieval, the word usually refers to an active process where a user enters a query consisting of one or more terms. In response the information retrieval system or search engine returns a (typically ranked) set of documents that correspond to that query.

Source: Herb Roitblat, Search 2020: The Glossary.

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Search Engine

A search component that implements the actual process of interpreting a search request and identifying subsets of documents. For example, a database management system such as Microsoft SQL Server contains a component that manages searches of the data stored in its databases.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Search Engine Optimization (SEO)

Changes made to a Web page that improves the positioning of that page with one or more search engines.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Search Engine Positioning (SEP)

The process of ordering Web sites or Web pages by a search engine or directory so that the most relevant sites appear first in the search results.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Search Hit

A document in ESI that is considered to match the requested Search Query.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Search Query

A well-formulated Search request that an automated search engine can interpret in order to produce matching results.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Search Results

A collection of Search Hits that match the intended documents of a Search Request.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Search Syntax

A particular search language required by a software program.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Search Term List

The list of keywords provided by a client for the purpose of searching a data set for responsive files using full text search.

Source: Ibis Consulting, Glossary.

Searchable TIFF

An imaged file accompanied, in a database, by OCR'd text that is searchable.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005).

See also:

GIF
Graphic Interchange File
Image file format
Joint photographic expert group
JPEG
Multi-page TIFF
PDF
PNG
Portable Document Format
Portable network graphic
Single-page TIFF
TIFF

SEC Regulation 10b(5)

Securities and Exchange Commission regulation governing the rights of shareholders. Many lawsuits by shareholders are filed under Rule 10b(5).

Source: Ibis Consulting, Glossary.

SEC Regulation 17a4

Securities and Exchange Commission regulation relating to data retention for financial services firms.

Source: Ibis Consulting, Glossary.

Second Request

A document request by either the Department of Justice (DOJ) or the Federal Trade Commission (FTC) for "additional information and documentary material relevant to [a] proposed acquisition" under the Hart-Scott-Rodino Antitrust Improvements Act of 1976 (the "HSR Act.")

Source: Ibis Consulting, Glossary.

Secure HTTP

See: S-HTTP (Secure HTTP)

Sedona / Sedona Conference

The Sedona Conference® (https://thesedonaconference.org) is a nonprofit, 501(c)(3) research and educational institute, founded in 1997 by Richard G. Braman, dedicated to the advanced study of law and policy in the areas of antitrust, complex litigation, and intellectual property rights. Sedona sponsors a preeminent think-tank in the area of Electronic Discovery known as Working Group 1 on Electronic Document Retention and Production. Sedona is well known for its thoughtful, balanced, and free publications, such as *The Sedona Conference® Glossary: E-Discovery & Digital Information Management* (Third Edition, Sept. 2010), *The Sedona Principles*

Addressing Electronic Document Production, Second Edition (June 2007), and *The Sedona Conference® Cooperation Proclamation* (July 2008).

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Seed Set

The initial Training Set provided to the learning Algorithm in an Active Learning process. The Documents in the Seed Set may be selected based on Random Sampling or Judgmental Sampling. Some commentators use the term more restrictively to refer only to Documents chosen using Judgmental Sampling. Other commentators use the term generally to mean any Training Set, including the final Training Set in Iterative Training, or the only Training Set in non-Iterative Training.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A collection of pre-categorized documents that is used as the initial training for a predictive coding system.

Source: Herb Roitblat, Predictive Coding Glossary.

Self Collection

A process where individual custodians identify and copy potentially relevant files for discovery.

Source: EDRM Collection Standards

Sender/Recipient Filter

A filter option that allows for including/excluding selected senders and recipients of e-mail messages.

Source: Ibis Consulting, Glossary.

See also:

Date filter	Filter
Extensions/sizes filter	MD5-known filter

Sensor

A mechanism for measuring some feature of the environment or device that can be used computationally.

Source: Herb Roitblat, Search 2020: The Glossary.

Sentiment Analysis

A process for identify the sentiment in a text (for example, a Tweet, blog post, or document). Typically sentiment analysis identifies whether the text expresses a positive (e.g., happiness) or negative (e.g., anger) emotion, though more subtle distinctions are also possible.

Source: Herb Roitblat, Search 2020: The Glossary.

SEO

See: Search Engine Optimization (SEO)

SEP

See: Search Engine Positioning (SEP)

Sequenced Packet Exchange (SPX)

A communications protocol used by Novell networks.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

External links:

Webopedia Computer Dictionary, http://www.webopedia.com/TERM/S/SPX.html.

Serial

Data stored or transmitted sequentially, as opposed to parallel.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Serial Line Internet Protocol (SLIP)

A communications standard used in Internet communications.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Serif

The little cross bars or curls at the end of strokes on type fonts. For example, in this sentence, the horizontal line at the bottom of the letter ‘r’.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Server

Any computer on a network that contains data or applications shared by users of the network on their client PCs.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Service Bureau

A vendor which performs ALS services such as photocopying, scanning, imaging, coding and, more recently, e-discovery services.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Service Level Agreement (SLA)

A service-level agreement is a contract that defines the technical support or business parameters that an application service provider or other IT outsourcing firm will provide its clients. The agreement typically spells out measures for performance and consequences for failure.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Service Pack

Software program bug fixes. They are available after release of the software program. They also usually address compatibility issues. Often, a certain service pack is required to enable certain software either to run or to run well.

SGML (Standard Generalized Markup Language)

An informal industry standard (lingua franca) for open systems document management which specifies the data encoding of a document's format and content.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A text based “language” for describing the content and structure of documents. SGML is used, for example, by some government agencies to publish reports that are useable by both machines and human readers. HTML is a simplified application of SGML.

See also:

HTML	JavaScript	XML
Java	SGML/HyTime	

SGML/HyTime

A multimedia extension to SGML, sponsored by DOD.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

HTML	JavaScript	XML
Java	SGML	

SHA-1

In cryptography, SHA-1 (Secure Hash Algorithm 1) is a cryptographic hash function designed by the United States National Security Agency and is a U.S. Federal Information Processing Standard published by the United States NIST. SHA-1 produces a 160-bit (20-byte) hash value known as a message digest. A SHA-1 hash value is typically rendered as a hexadecimal number, 40 digits long.

Source: Wikipedia, SHA-1, https://en.wikipedia.org/wiki/SHA-1

See also:

Hash	Hashing / Hash / Hash Value	MD5
Hash value		

Shadow IT

Projects, devices, or software that are used by employees without the control, permission, or even the awareness of the organization's information technology program. Shadow IT is related to BYOD (bring your own device), but shadow IT implies an ungoverned use of technology, and BYOD may have limited or even deep controls.

Source: Herb Roitblat, Search 2020: The Glossary.

Shingling

A Feature Engineering method in which the Features consist of all N-Grams in a text, for some number N. For example, the Trigram Shingling of the text "To be or not to be" consists of the Features "to be or"; "be or not"; "or not to"; and "not to be." Note that the Features overlap one another in the text, suggesting the metaphor of roof shingles.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Sibling

A sibling is a document that shares a common parent with the document in question (e.g. two attachments that share the same parent email or are sibling documents in the same Zip file).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Signal Detection Theory

Invented at the same time and in conjunction with radar, the science of distinguishing true observations from spurious ones. Signal Detection Theory is widely used in radio engineering and medical diagnostic testing. The terms True Positive, True Negative, False Positive, False

Negative, Sensitivity, Specificity, Receiver Operating Characteristic Curve, Area Under the ROC Curve, and Internal Response Curve, all arise from Signal Detection Theory.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Significance / Significant

The confirmation, with a given Confidence Level, of a prior hypothesis, using a Statistical Estimate. The result is said to be Statistically Significant if all values within the Confidence Interval for the desired Confidence Level (typically 95%) are consistent with the hypothesis being true, and inconsistent with it being false. For example, if the hypothesis is that fewer than 300,000 Documents are Relevant, and a Statistical Estimate shows that, 290,000 Documents are Relevant, plus or minus 5,000 Documents, we say that the result is Significant. On the other hand, if the Statistical Estimate shows that 290,000 Documents are Relevant, plus or minus 15,000 Documents, we say that the result is not Significant, because the Confidence Interval includes values (i.e., the values between 300,000 and 305,000) that contradict the hypothesis.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Statistically significant means that the observed results are unlikely to have occurred by chance. Used in statistical decisions to decide whether a difference, for example, is large enough that it is unlikely to happened by chance from the sampling distribution. In statistics, in general, significance, refers to whether the outcome is so unlikely under the null hypothesis (no real difference) that we reject the null hypothesis and accept the alternative. For example, we select a random sample of students from each of two schools and we measure their reading comprehension. The null hypothesis is that there is no difference between schools on reading comprehension. The motivated hypothesis is that there is a difference. If the difference between mean (average) reading comprehension of these two samples is sufficiently large that it is unlikely, then we say that the difference is significant, and that the two schools differ in their reading comprehension. It is a misnomer to speak about a significant random sample. Significance refers to this kind of hypothesis test, not the size of the sample.

Source: Herb Roitblat, Search 2020: The Glossary.

Source: Herb Roitblat, Predictive Coding Glossary.

Similar Document Search

A search that finds all documents similar to the primary document.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Ad Hoc Search

Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

SIMM (Single, In-Line Memory Module)

A search that finds all documents similar to the primary document.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Simple Mail Transfer Protocol

See: SMTP (Simple Mail Transfer Protocol)

Simplex

One-sided page(s).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Simplex Scanner

A document scanner that copies single-sided documents.

Source: RSI, Glossary.

See also:

Double-sided scanner
Duplex scanner
Flatbed scanner
Scanner

Single, In-Line Memory Module

See: SIMM (Single, In-Line Memory Module)

Single-Mail Archive

An aggregate of e-mail messages and attachments saved outside of multi-mail containers like PST or NSF. Single-mail archives can take the form of MSG, EML, TXT, HTML, or other file formats.

Source: Ibis Consulting, Glossary.

See also:

Container
EML
Mail container
Mailbox
MSG
Multi-mail container
NSF
OST
PST
RFC compliant email
RFC822
Single-mail container
SMTP

Single-Mail Container

A container, or file, that holds a single e-mail message, such as EML and MSG.

See also:

Container
EML
Mail container
Mailbox
MSG
Multi-mail container
NSF
OST
PST
RFC compliant email
RFC822
Single-mail archive
SMTP

Single-Page Text

Extracted, single-page text files.

Source: Ibis Consulting, Glossary.

Single-Page TIFF

The standard output format for TIFF images, where one page = one TIFF image. Some load file formats require multiple TIFF files merged into one TIFF file with multiple pages.

See also:

GIF
Graphic Interchange File
Image file format
Joint photographic expert group
JPEG
Multi-page TIFF
PDF
PNG
Portable Document Format
Portable network graphic
Searchable TIFF
TIFF

Singular Value Decomposition

A mathematical technique that summarizes the correlation between items and their features. One of the techniques used in eDiscovery as the basis for concept search, where the items are documents and the features are words.

Source: Herb Roitblat, Search 2020: The Glossary.

Skew

During printing or scanning, the contents of a page are almost never exactly vertical, which referred to as being skewed.

Source: RSI, Glossary.

See also:

De-Skew

SLA

See: Service Level Agreement (SLA)

Slack Space

A form of residual data, slack space is the amount of on-disk file space from the end of the logical record information to the end of the physical disk record. Slack space can contain information soft-deleted from the record, information from prior records stored at the same physical location as current records, metadata fragments and other information useful for forensic analysis of computer systems.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Remnant data from deleted files still located in clusters on a hard drive.

Source: RenewData, Glossary (10/5/2005).

The difference in empty bytes of the space that is allocated in clusters minus the actual size of the files. Also described as the data fragments stored randomly on a hard drive during the normal operation of a computer, or the residual data left on the hard drive after new data has overwritten some of the previously stored data.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

The difference between the size of a file and the size of the various clusters where it is stored, since the file segments may be smaller than the clusters where they reside. May also refer to data fragments stored randomly on a hard drive during the normal operation of a computer or residual data left on a hard drive after new data has overwritten deleted files.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ambient data	Free space	Swap file
Fragmented data	Residual data	Unallocated space

SLIP

See: Serial Line Internet Protocol (SLIP)

Small Computer Systems Interface

See: SCSI (Small Computer Systems Interface)

Smart Card

A credit card size device which contains a microprocessor, memory and a battery.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

SMP (Symmetric Multi-Processing)

A system design of multiple CPUs in which any CPU can be assigned any application task. Typically, one CPU is the controller and handles system boot, I/O requests, and distribution of tasks to the other CPUs.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

SMTP (Simple Mail Transfer Protocol)

Simple Mail Transfer Protocol (SMTP) is an Internet standard for electronic mail (email) transmission. First defined by RFC 821 in 1982, it was last updated in 2008 with the Extended SMTP additions by RFC 5321—which is the protocol in widespread use today.

Source: Wikipedia, Simple Mail Transfer Protocol, https://en.wikipedia.org/wiki/Simple_Mail_Transfer_Protocol

(pronounced as separate letters) Short for Simple Mail Transfer Protocol, a protocol for sending e-mail messages between servers. Most e-mail systems that send mail over the Internet use SMTP to send messages from one server to another; the messages can then be retrieved with an e-mail client using either POP or IMAP. In addition, SMTP is generally used to send messages from a mail client to a mail server. This is why you need to specify both the POP or IMAP server and the SMTP server when you configure your e-mail application.

Source: Webopedia, SMTP - Simple Mail Transfer Protocol definition, http://www.webopedia.com/TERM/S/SMTP.html

See also:

Container	Multi-mail container	RFC822
EML	NSF	Single-mail archive
Mail container	OST	Single-mail container
Mailbox	PST	
MSG	RFC compliant email	

Social Network Analysis

Investigations of who in an organization is communicating with whom. These connections are often displayed as a network diagram, with individuals as nodes and the emails or other communications between them as links. Social networks are often useful to determine how information has been flowing through an organization. They can also help to identify individuals with specific kinds of knowledge.

Source: Herb Roitblat, Search 2020: The Glossary.

Software

Any set of instructions stored on computer-readable media that tells a computer what to do. Includes operating systems and software applications.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: RSI, Glossary.

A series of files containing instructions to the computer for performing functions. A software "program" contains the instructions to accept data in certain formats.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Coded instructions (programs) that make a computer do useful work.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Software Application

See: Application

Sort

Putting a report in a particular order, such as chronological or numerical.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Sound-Alike

A search method whereby the computer produces a list of words that "sound" similar to the desired word and can themselves be searched.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Stemming
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Source File

The raw data received from a client, either on digital media or uploaded to the network.

Source: Ibis Consulting, Glossary.

Splatter

Data that should be kept on one disc of a jukebox goes instead to multiple platters.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Spoliation

Spoliation is the destruction or alteration of evidence during on-going litigation or during an investigation or when either might occur sometime in the future. Failure to preserve data that may become evidence is also spoliation.

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Norcross Group FAQ's, http://norcrossgroup.com/faq.html#14.

Generally, the intentional or negligent destruction or alteration of evidence when there is current litigation or an investigation or there is reasonable anticipation that either may occur in the near future. Some jurisdictions also define it as a failure to preserve information that may become evidence.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

The intentional alteration or destruction of a relevant document or documents.

Source: Ibis Consulting, Glossary.

The original legal definition was the destruction of a thing by the act of a stranger; as in the erasure or alteration of a writing by the act of a stranger. In e¬discovery cases the focus has been on the intentional nature of the act, which can include deletion, partial destruction or alteration, generally by a party to the action or someone under their control.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Spoliation is the destruction of records which may be relevant to ongoing or anticipated litigation, government investigation or audit. Courts differ in their interpretation of the level of intent required before sanctions may be warranted.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The intentional, negligent, or reckless, loss, destruction, alteration or obstruction of relevant evidence.

SPP (Standard Parallel Port)

A standard parallel port (SPP) is a port for connecting various relatively high bandwidth peripherals, most commonly printers, to a PC. Later versions of the SPP allow duplex communication. They use the DB-25 connector. The original SPP, by Centronics, was introduced in 1970 and soon became the de facto industry standard. However, a number of different manufacturers used the SPP with a variety of connectors, such as the DC-35, the DD50 and the M50.

Source: Techopedia, Standard Parallel Port (SPP), https://www.techopedia.com/definition/3667/standard-parallel-port-spp

Spreadsheet

A software program that arranges data in a matrix of cells and performs calculations based on the arrangement of the cells. The most popular spreadsheets are Lotus 1-2-3 and Microsoft Excel.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A compilation of data in table form arranged in columns and rows. Programs such as Excel build files that contain one or more spreadsheets. Spreadsheets may also contain graphic and other elements and may hidden data.

Spreadsheet Formula

A mathematical formula applied to a cell to calculate the contents, e.g. cell A3= cell A 1+ cell A2.

SPX

See: Sequenced Packet Exchange (SPX)

SQL (Structured Query Language)

A type of relational database management system (RDBMS). Relationships in a relational database are represented by linkages that exist between two or more pieces of data. The final defining feature of SQL is its ability to return data from one data field based on its relationship with another data field. See also Relational Database Management Systems.

Source: EDRM Search Glossary.

SQL is a standard programming language for getting information from and updating a database. Although SQL is a standard, many database products support SQL with proprietary extensions to the standard language.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The popular standard for running database searches (queries) and reports.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Structured Query Language, the language used to control traditional relational databases. A relational database stores data in one or more tables. It can also represent relations between the columns of one table (variables in that table) and columns (variables) in another table, hence the name “relational database.”

Source: Herb Roitblat, Search 2020: The Glossary.

See also:

Database
Flat file database
Full text database
Relational database
WAIS - wide area information server

Stand Alone Computer

A single computer not connected to a network.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A single computer, as distinct from a computer attached to a network.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A personal computer that is not connected to any other computer or network, except possibly through a modem.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

Client/server network	Network	WAN - wide area network
LAN - local area network	Peer-to-peer network	
MAN - metropolitan area network	SAN - storage area network	

Standard Generalized Markup Language

See: SGML (Standard Generalized Markup Language)

Standard Parallel Port

See: SPP (Standard Parallel Port)

Statistical Estimate

A quantitative estimate of a Population characteristic using Statistical Estimation. It is generally expressed as a Point Estimate accompanied by a Margin of Error and a Confidence Level, or as a Confidence Interval accompanied by a Confidence Level.

Source; Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Statistical Estimation

The act of estimating the Proportion of a Document Population that has a particular characteristic, based on the Proportion of a Random Sample that has the same characteristic. Methods of Statistical Estimation include Binomial Estimation and Gaussian Estimation.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Statistical Model

A mathematical abstraction of the Document Population that removes irrelevant characteristics while largely preserving those of interest for a particular purpose. For the purpose of computing Recall, a Statistical Model need only consider whether or not the Documents are Relevant, and whether or not the Documents are Coded Relevant, not any other characteristics of the Documents.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Statistical Sample / Statistical Sampling

A method in which a Sample of the Document Population is drawn at random, so that statistical properties of the Sample may be extrapolated to the entire Document Population; the Sample resulting from such action.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Status

A common element within each e-discovery Phase which refers to the activities, tasks and methods undertaken in relation to a defined objective within a Phase. In Project Management, benchmarking of current work against expressed, intended or expected outcome, and reporting on same.

Source: EDRM Metrics Glossary

Stemming

A search option that returns matches for all variations of the root word of the initial query word. For example, if the query word was sing, then if a search used stemming the search results would match singing, sang, sung, song, and songs as well as sing.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

In Keyword or Boolean Search, or Feature Engineering, the process of equating all forms of the same root word. For example, the words “stem,” “stemming,” “stemmed,” and “stemmable” would all be treated as equivalent, and would each yield the same result when used as a Search Terms in a Query. In some search systems, stemming is implicit and in others, it must be made explicit through particular Query syntax.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The process of removing prefixes and suffixes from words before indexing them and as part of query processing. For example, the word “swimming” could be stemmed to “swim.” If words are stemmed as they are indexed, the query must also stem the words so that the query can match the index. In a system that uses stemming, several word forms can be indexed identically, for example, “swimmer” and “swimming” would both be indexed as “swim.”

Source: Herb Roitblat, Search 2020: The Glossary.

See also:

Ad Hoc Search

Adaptive pattern recognition

Associative retrieval

Boolean search

Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Synonym search
Term search
Topical search
Weighted relevance search
Wildcard search

Stipulation

Stipulation is an agreement made between opposing parties prior to a pending hearing or trial. For example, both parties might stipulate to certain facts, and therefore not have to argue those facts in court. After the stipulation is entered into, it is presented to the judge.

Source: EDRM Presentation Guide.

Stop Word

A common word that is eliminated from Indexing. Eliminating Stop Words from Indexing dramatically reduces the size of the Index, while only marginally affecting the search process in most circumstances. Examples of Stop Words include “a,” “the,” “of,” “but,” and “not.” Because phrases and names such as “To be or not to be,” and “The Who,” contain exclusively Stop Words that would not be Indexed, they would not be identified (or identifiable) through a Keyword Search.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Words that are so commonly used that there is little loss by ignoring them in a query or when the documents are indexed. Many systems involve stopword lists, which may include words like “I”, “he”, “are”, and “is.”

Source: Herb Roitblat, Search 2020: The Glossary.

Storage Area Network

See: SAN (Storage Area Network)

Storage Device

Any device that a computer uses to store information.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Disk drive	Magneto-optical drive	Zip drive
Floppy disk drive	Portable drive	
Jaz drive	Tape drive	

Storage Media

Any removable device that stores data. See magnetic or optical storage media.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

CD	DVD	Magnetic disk
CD-R	DVD-ROM	Magnetic storage media
CD-ROM	Floppy disk	Media
CD-RW	Hard disk	Optical disk
Disc	Hard drive	WORM disk
Disk	Jaz disk	Zip disk
Diskette	Laser disc	

Store

To place information onto a disk where it is available for later use.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Stratified Sampling

A form of random sampling in which the population is formed into subgroups or “strata.” Objects in each group are sampled in the same proportion as the size of the group is to the whole population. Each object has an equal chance of being sampled, but it also ensures that each group is sampled proportionately.

Source: Herb Roitblat, Predictive Coding Glossary.

Structured Data

Structured Data is data that is organized. The most common type is database content. It refers to any type of data organized such as Internet data or other types of data that has been indexed.

Source: EDRM Metrics Glossary

Data that resides in a fixed field within a record or file is called structured data. This includes data contained in relational databases and spreadsheets.

Source: http://www.webopedia.com/TERM/S/structured_data.html.

Structured Query Language

See: SQL (Structured Query Language)

Subject Category

A data field in a database used to capture specific subject codes.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Summary
Text

Subject Code

A code for a case-specific legal or factual subject.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Subject Matter Expert(s)

One or more individuals (typically, but not necessarily, attorneys) who are familiar with the Information Need and can render an authoritative determination as to whether a Document is Relevant or not.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Subjective Coding

The Subjective Coding of a document involves linking a legal interpretation to an individual document. In direct opposition to objective coding, in which bibliographic data about the document is recorded. Subjective Coding types include the classification of documents as privileged and responsive, and the categorization of documents by legal issue ("issue coding").

Source: EDRM Metrics Glossary

Entering information from a document that requires the coder to exercise judgment, such as subject or issue codes. This field is often left blank for the law firm's paralegals or associates to fill in.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The coding of a document using legal interpretation as the data that fills a field. Performed by paralegals or other trained legal personnel.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Categorizing documents by their responsiveness to specific case issues or topics.

See also:

Bibliographic Coding	Issue coding	Taxonomic coding
Coding	Level coding	Verbatim coding
Indexing	Objective coding	
Issue Code	Tag	

Subtractive Colors

Since the colors of objects are white light minus the color absorbed by the object, they are called subtractive. This is how ink on paper works. The subtractive colors of process ink are CMYK (Cyan, Magenta, Yellow and Black) and are specifically balanced to match additive colors (RGB).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Summary

A data field in a database that records the summary of a document.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field
Cross-reference field
Customized data field
Customized field definition
Data field definition
Date field
End document number
Field
Index/coding field
Key field
Marginalia
Names mentioned in text
Note field
Other number field
Production source
Recipient
Subject category
Text

Super Video Graphic Adapter (SVGA)

A video graphic adapter which exceeds the minimum VGA standard of 640 by 480 by 16 colors. Can reach 1600 by 1280 and 256 colors.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Supervised Learning

A Machine Learning method in which the learning Algorithm infers how to distinguish between Relevant and Non-Relevant Documents using a Training Set. Supervised Learning can be a stand-alone process, or used repeatedly in an Active Learning process.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A kind of machine learning where the objects are labeled by an exterior source, typically, a subject matter expert. The goal of supervised learning is typically to replicate the decision pattern of the outside expert and apply the same patterns to previously unseen objects.

Source: Herb Roitblat, Search 2020: The Glossary.

Support Vector Machine

A state-of-the-art Supervised Learning Algorithm that separates Relevant from Non-Relevant Documents using geometric methods (i.e., geometry). Each Document is considered to be a point in [hyper]space, whose coordinates are determined from the Features contained in the Document. The Support Vector Machine finds a [hyper]plane that best separates Relevant from Non-Relevant Training Examples. Documents outside the Training Set (i.e., uncoded Documents from the Document Collection) are then Classified as Relevant or not, depending on which side

of the [hyper]plane they fall on. Although a Support Vector Machine does not calculate a Probability of Relevance, one may infer that the Classification of Documents closer to the [hyper]plane is less certain than for those that are far from the [hyper]plane.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A machine-learning approach, used for categorizing data. The goal of the SVM is to learn the boundaries that separate two or more classes of objects. Given a set of already categorized training examples, an SVM training algorithm identifies the differences between the examples of each training category and can then apply similar criteria to distinguishing future examples.

Source: Herb Roitblat, Search 2020: The Glossary.

Source: Herb Roitblat, Predictive Coding Glossary.

SVGA

See: Super Video Graphic Adapter (SVGA)

Swap File

A swap file (or swap space or, in Windows NT, a pagefile) is a space on a hard disk used as the virtual memory extension of a computer's real memory (RAM). Having a swap file allows your computer's operating system to pretend that you have more RAM than you actually do. The least recently used files in RAM can be "swapped out" to your hard disk until they are needed later so that new files can be "swapped in" to RAM. In larger operating systems (such as IBM's OS/390), the units that are moved are called pages and the swapping is called paging.

Source: TechTarget, swap file (swap space or pageful) definition, http://searchwindowsserver.techtarget.com/definition/swap-file-swap-space-or-pagefile

See also:

Ambient data	Free space	Slack space
Fragmented data	Residual data	Unallocated space

Symmetric Multi-Processing

See: SMP (Symmetric Multi-Processing)

Synonym Search

A synonym search returns documents that contain terms similar in meaning to the query words, usually using a thesaurus to determine which terms would match the query words.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Term search
Topical search
Weighted relevance search
Wildcard search

Synonymy

Having the equivalence of meaning; having the same definition without having the same expression.

Source: EDRM Search Glossary.

Synthetic Document

An industry-specific term generally used to describe an artificial Document created by either the requesting party or the producing party, as part of a Technology-Assisted Review process, for use as a Training Example for a Machine Learning Algorithm. Synthetic Documents are contrived Documents in which one party imagines what the evidence might look like and relies on the Machine Learning Algorithm to find actual Documents that are similar to the artificial Document.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Sysadmin

A system administrator, or sysadmin, is a person who is responsible for the upkeep, configuration, and reliable operation of computer systems; especially multi-user computers, such as servers.

Source: http://en.wikipedia.org/wiki/System_administrator.

The person in charge of keeping a network working. Also referred to as *sysadmin* or *sysop*.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Sysop

See: Sysadmin

System Administrator

See: Sysadmin

System Program

Programs that control the internal operations of a computer system. Examples are operating systems, compilers, interpreters, assemblers, and mathematical routines.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

System Registry

The system configuration files used by Microsoft Windows to store settings about user preferences, installed software, hardware and drivers and other settings required for Windows to run correctly.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Systematic Sample / Systematic Sampling

A Sampling method in which every Nth Document (for some fixed number N) is selected, when the Documents are considered in some prescribed order; the Sample resulting from such action. A Systematic Sample is random (and hence a true Statistical Sample) only when the prescribed order is itself random. Sometimes referred to as an Interval Sample / Interval Sampling.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Systems

A common element within each e-discovery Phase which refers to computer storage devices, active applications for the storage or use of data or ESI; or to work processes designed to achieve a specified result.

Source: EDRM Metrics Glossary

T

Tag

An emendation added to a document during the review process. Tags can be used to assign documents to issues, to indicate which ones should be printed, or for any other reason the case requires.

Tagged Image File Format (TIFF)

A graphic file format used for storing still-image bitmaps. TIFFs are stored in tagged fields, and programs use the tags to accept or ignore fields, depending on the application.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Fenwick & West LLP, FWPS eDiscovery Terminology (11/6/2005). Citing Fios' eDiscovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary_sz.html.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

A widely used bit-mapped graphics file format. This is essentially a picture of a document.

Source: RenewData, Glossary (10/5/2005).

A bit mapped graphics file format that contains a picture of a document.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Graphic files that portray a single page of a file for viewing purposes with a .tif extension (in the case of Multi-page TIFFs, output images can consist of multiple pages).

Source: Ibis Consulting, Glossary.

One of several standards for making electronic images.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The "de facto" electronic/computer standard for scanned, bit-mapped images – 8 bit color and gray scale. Originated in 1986 as a joint project of Microsoft and Aldus. Includes several types and groups which are compressed & uncompressed.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

One of the most widely supported file formats for storing bit-mapped images. Files in TIFF format often end with a .tif extension. 10

This image format is commonly used as the standard file delivery format for production.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

See also:

GIF

Graphic Interchange File

Image file format

Joint photographic expert group

JPEG

Multi-page TIFF

PDF

PNG

Portable Document Format

Portable network graphic	Searchable TIFF	Single-page TIFF

Tape

Random memory which can be read but not written (i.e. changed).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Backup	Digital audio tape	Media
Backup tape	Disaster recovery tape	QIC - quarter inch cartridge
DAT - digital audio tape	DLT - digital linear tape	
Data extraction	Magnetic storage media	

Tape Backup

A process of copying electronic data from a storage device, such as a computer's hard drive, to a tape cartridge device. This security measure ensures that the data is not lost in the event of an equipment failure or disaster. Tape backup can be achieved manually or programmed to occur automatically.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Tape Drive

A hardware device used to store data on a magnetic tape. Tape drives are usually used to back up large quantities of data due to their large capacity and cheap cost relative to other data storage options.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Disk drive	Magneto-optical drive	Zip drive
Floppy disk drive	Portable drive	
Jaz drive	Storage device	

TAPI (Telephony Application Programming Interface)

A Microsoft-based standard for basic telephone services that allows a PC to access phone books, control phone equipment, and interface with voice-mail and e-mail systems.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

TAR

See: Technology-Assisted Review (TAR)

Targa Image File Format (TGA)

This is a "scanned format" – widely used for color-scanned materials (24-bit) as well as by various "paint" and desktop publishing packages

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Targeted Collection Strategies

A targeted collection strategy is one that is specifically designed to avoid over collection of data that is known to be irrelevant or potentially irrelevant. A non-targeted strategy is designed to collect all the data from a particular storage device or repository in a comprehensive manner. Culling and filtering protocols are subsequently applied to the data corpus to either eliminate non-responsive data or isolate responsive data. A targeted strategy takes into consideration culling and filtering protocols at the point of collection (e.g. only collecting a custodian's email inbox as opposed to imaging their entire hard drive).

Source: EDRM Metrics Glossary

Taxonomic Coding

See also:

Bibliographic Coding
Coding
Indexing
Issue Code
Issue coding
Level coding
Objective coding
Subjective coding
Tag
Verbatim coding

Taxonomy

A hierarchical organizational scheme that arranges the meanings of words into classes and subclasses. For example, vehicles, aircraft, and ships are modes of transportation; cars, trucks, and bicycles are vehicles, and Fords and Chryslers are cars.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A specific coding language and terminology developed for use in a particular case.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A hierarchical categorical structure where each class may contain one or more subclasses. The scientific classification system of organizing plants and animals into a specific phylum, family, genus, species, etc. is an example of a taxonomic system. Each subclass is an example of its parent class.

Source: Herb Roitblat, Search 2020: The Glossary.

TB (Terabyte)

A trillion bytes, or a million megabytes. The entire collection of the Library of Congress would equal approximately 20 terabytes if digitized.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A trillion bytes, or more correctly 1,024 megabytes.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

A terabyte is a measure of computer data storage capacity and is one thousand billion (1,000,000,000,000) bytes.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

A 1000 Gigabytes (GB) or 1,099,511,627,776 bytes.

See also:

Bit	MB - megabyte	EB - exabyte
Byte	GB - gigabyte	
KB - kilobyte	PB - petabyte	

TCP (Transmission Control Protocol)

The protocol used in conjunction with Internet Protocol (IP) to transmit information over the Internet in the form of units.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

TCP/IP (Transmission Control Protocol/Internet Protocol)

A collection of protocols that define the basic workings of the features of the Internet.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Network communications protocol. This is the protocol used by the Internet.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Technology-Assisted Review (TAR)

A process for Prioritizing or Coding a Collection of Documents using a computerized system that harnesses human judgments of one or more Subject Matter Expert(s) on a smaller set of documents and then extrapolates those judgments to the remaining Document Collection. Some TAR methods use Machine Learning Algorithms to distinguish Relevant from Non-Relevant Documents, based on Training Examples Coded as Relevant or Non-Relevant by the Subject Matter Experts(s), while other TAR methods derive systematic Rules that emulate the expert(s)' decision-making process. TAR processes generally incorporate Statistical Models and/or Sampling techniques to guide the process and to measure overall system effectiveness.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Any of a number of technologies that use technology, usually computer technology, to facilitate the review of documents for discovery.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

CAR Predictive Coding TAR

Telecommunications

Data transmission between a computer system and remote devices, usually over telephone lines.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Telephony

Converting sounds into electronic signals for transmission.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Telephony Application Programming Interface

See: TAPI (Telephony Application Programming Interface)

Terabyte

See: TB (Terabyte)

Term Frequency and Inverse Document Frequency (TF-IDF)

An enhancement to the Bag of Words method in which each word has a weight based on Term Frequency – the number of times the word appears in the Document – and Inverse Document Frequency – reciprocal of the number of Documents in which the word occurs.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Term Search

A variant of keyword search, with the emphasis on searching for combinations of words such as phrases.

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Topical search
Weighted relevance search
Wildcard search

Terminal

A device with input and output devices (keyboard and monitor) connected to a computer system.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Text

A data field that allows the entry of text in a manner similar to word processing software, but is limited to a specific number of characters. Text fields can be sorted and are typically used for names.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Attachment field
Attorney notes field
Author field
Beginning document number
Beginning number field
Copyee field

Cross-reference field	Field	Other number field
Customized data field	Index/coding field	Production source
Customized field definition	Key field	Recipient
Data field definition	Marginalia	Subject category
Date field	Names mentioned in text	Summary
End document number	Note field	

Text Clustering

Text clustering is a technology that analyzes a document collection and organizes the documents into groups based on finding documents that are similar to each other based on words contained within it (such as noun phrases). Text clustering establishes a notion of "distance between documents" and attempts to select enough documents into the cluster so as to minimize the overall pair-wise distance among all pairs of documents.

Source: EDRM Search Glossary.

Text Extraction

The process of pulling the text and data from electronic documents for the purposes of loading the data into a database. The process removes formatting, and graphics from a document leaving only the text.

Text REtrieval Conference

See: TREC

TF-IDF

See: Term Frequency and Inverse Document Frequency (TF-IDF)

TF-IDF

In information retrieval, a weighting procedure so that some words in a query or document get emphasized more than others. A document is ranked higher using TF-IDF when it has more occurrences of the query term (TF or term frequency) and ranks lower when the word occurs in more documents (IDF or inverse document frequency). There are different rules for deciding how to combine TF with IDF, on common rule is to rank the documents based on the ratio of TF to log(IDF).

Source: Herb Roitblat, Search 2020: The Glossary.

TGA

See: Targa Image File Format (TGA)

The Generally Accepted Recordkeeping Principles ® ("The Principles")

The Principles reflect standards and guidelines related to records management, developed by ARMA International, a not-for-profit professional association and a widely-recognized authority on managing records and information. The Principles include: (1) Principle of Accountability, (2) Principle of Integrity, (3) Principle of Protection, (4) Principle of Compliance, (5) Principle of Availability, (6) Principle of Retention, (7) Principle of Disposition, and (8) Principle of Transparency.

Source: IGRM White Paper

Thesaurus Expansion

In Keyword or Boolean Search, replacing a single Search Term by a list of its synonyms, as listed in a thesaurus.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Thing

A thing in the context of Internet of things (IoT), is any object that could be connected to the Internet, each of which would have a unique URI.

Source: Herb Roitblat, Search 2020: The Glossary.

Threading

Organizing emails into conversational groups. For example, if John sends an email to Mary and she replies, both emails are part of the same conversational thread.

Source: Herb Roitblat, Search 2020: The Glossary.

Thumb Drive

Also known as keychain drive, thumb drive and USB flash drive.

Thumbnail

A small version of an image used for quick overviews or to get a general idea of what the image looks like.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

TIFF

See: Tagged Image File Format (TIFF)

TIFF Group III

A one-dimensional compression format for storing black and white images that is utilized by most fax machines.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

TIFF Group IV

A two-dimensional compression format for storing black and white images. Typically compresses at a 20-to-1 ratio for standard business documents.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

TIFFing

The process of opening files in their native applications, extracting text and printing them as TIFF images.

Source: Ibis Consulting, Glossary.

Time

Refers to the measurable hours involved in each identifiable task, activity or action. Time is also a variable element of Cost and Volume.

Source: EDRM Metrics Glossary

Tokenization

An operation that examines a document or block of text and breaks the text into words. Typically, a space is used to separate words, but special characters such as a hyphen, period, or quotation mark can also be used.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Tool Kit Without An Interesting Name (TWAIN)

A universal toolkit with standard hardware/software drivers for multi-media peripheral devices.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Topical Search

A vertical search engine as distinct from a general web search engine, focuses on a specific segment of online content. They are also called specialty or topical search engines. The vertical content area may be based on topicality, media type, or genre of content. Common verticals include shopping, the automotive industry, legal information, medical information, scholarly

literature, and travel. Examples of vertical search engines include; Mocavo, Nuroa, Trulia and Yelp. In contrast to general web search engines, which attempt to index large portions of the World Wide Web using a web crawler, vertical search engines typically use a focused crawler which attempts to index only relevant web pages to a pre-defined topic or set of topics.

Source: Wikipedia, Vertical search, https://en.wikipedia.org/wiki/Vertical_search

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Weighted relevance search
Wildcard search

Training Example

One Document from a Training Set.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Training Set

A Sample of Documents Coded by one or more Subject Matter Expert(s) as Relevant or Non-Relevant, from which a Machine Learning Algorithm then infers how to distinguish between Relevant and Non-Relevant Documents beyond those in the Training Set.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Transmission Control Protocol

See: TCP (Transmission Control Protocol)

Transmission Control Protocol/Internet Protocol

See: TCP/IP (Transmission Control Protocol/Internet Protocol)

Transmission Speed

The rate at which data passes through communications lines; usually measured in bits per second (bps).

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

TREC

The Text REtrieval Conference, sponsored by the National Institute of Standards and Technology (NIST), which has run since 1992, to "support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. In particular, the TREC workshop series has the following goals: to encourage research in information retrieval based on large test Collections; to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas; to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems."

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

The Text REtrieval Conference, organized the US National Institute of Standards and Technology. TREC is an annual conference and friendly competition among information retrieval systems intended to promote the science of text retrieval. For several years TREC included a legal track, which investigated text retrieval in the context of discovery.

Source: Herb Roitblat, Search 2020: The Glossary.

External link:

Text REtrieval Conference (TREC), http://trec.nist.gov

TREC Legal Track

From 2006 through 2011, TREC included a Legal Track, which sought "to assess the ability of information retrieval techniques to meet the needs of the legal profession for tools and methods capable of helping with the retrieval of electronic business records, principally for use as evidence in civil litigation."

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Triage

The process of reviewing a document set (discovery set) for responsiveness and/or privilege. Triage refers to the practice of quickly identifying which documents require additional attention and which can be easily classified as either responsive or nonresponsive.

Trigram

An N-Gram where N = 3 (i.e., a 3-gram).

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

True Negative (TN)

A Non-Relevant Document that is correctly identified as Non-Relevant by a search or review effort.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

One of four response states in a categorization task. True negative responses are those that are truly in the negative category and are classified as negative.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

False Negative (FN) False Positive (FP) True Positive (TP)

True Negative Rate (TNR)

The fraction (or Proportion) of Non-Relevant Documents that are correctly identified as Non-Relevant by a search or review effort.

Source; Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

True Positive (TP)

A Relevant Document that is correctly identified as Relevant by a search or review effort.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

One of four response states in a categorization task. True positive responses are those that are truly in the positive category and are classified as positive.

Source: Herb Roitblat, Predictive Coding Glossary.

See also:

False Negative (FN) False Positive (FP) True Negative (TN)

True Positive Rate (TPR)

The fraction (or Proportion) of Relevant Documents that are correctly identified as Relevant by a search or review effort. True Positive Rate is a term used in Signal Detection Theory; Recall is the equivalent term in Information Retrieval.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

True Resolution

The "true" optical resolution of a scanner is the number of pixels per inch (without any software enhancements).

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Truncation

A Search Specification that indicates that matching documents must contain words that begin with the letters entered, but that the matching words can end with any combination of letters.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

TWAIN

See: Tool Kit Without An Interesting Name (TWAIN)

TWAIN Scanner Driver

A specialized application used for communication between scanners and computers.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Two-Tailed Distribution

See: Two-Tailed Test

Two-Tailed Test

A confidence interval that is arranged symmetrically around the average or mean of a distribution. The tails, outside of the confidence interval are of equal size. Also called two-tailed distribution.

Source: Herb Roitblat, Predictive Coding Glossary.

Typeface

There are over 10,000 typefaces available for computers. The general categories are:

1. Oldstyle: Faces have slanted serifs, gradual thick to thin strokes and a slanted stress (the "O" appears slanted)
2. Modern: Faces have thin, horizontal serifs, radical thick to thin strokes and a vertical street (the "O" does not appear to slant.)
3. Slab Serif: Faces have thick, horizontal serifs, little or no thick-to-thin in the strokes and a vertical stress (the "O" appears vertical).
4. Sans Serif: Faces have no serifs.
5. Script: From elaborate handwriting styles to casual, freeform, unconnected letter forms.
6. Decorative: Unusual fonts, designed to be very different and attention getting.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

U

Ultrafiche

Microfiche which can hold 1,000 documents/sheet as opposed to the normal 270.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Unallocated Space

Space on a hard drive that potentially contains intact files, remnants of files, subdirectories, or temporary files that were created and then deleted by either a computer application, the operating system or the operator.

Source: RenewData, Glossary (10/5/2005).

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

See also:

Ambient data	Free space	Slack space
Fragmented data	Residual data	Swap file

Uncertainty Sampling

An Active Learning approach in which the Machine Learning Algorithm selects the Documents as to which it is least certain about Relevance, for Coding by the Subject Matter Expert(s), and addition to the Training Set.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Unicode / Unicode Transformation Format

All electronic data is represented as sequences of bits, or numbers. Each alphabet or script used in a language is mapped to a unique numeric value, or 'encoded' for use on a computer using a standard known as Unicode. Within Unicode, each letter or character has been assigned its own unique value in the Unicode encoding schemes, known as the Unicode Transformation Format (UTF). The UTF utilizes multiple encoding schemes, of which the most commonly used are known as UTF-8 and UTF-16. For example, the English alphabet and the more common punctuation marks have been assigned values between 0 and 255, while Tibetan characters have been assigned the values between 3,840 (written as x0F00) and 4,095 (written as x0FFF). All modern (and many historical) scripts are supported by the Unicode Standard. Unicode provides a unique number for every character, regardless of the platform, program, or language. The Unicode Standard is described in detail at the website http://www.unicode.org. See also, Character Encoding.

Source: EDRM Search Glossary.

Unified Governance

Unified Governance is a marriage between policy integration and process transparency. Effective unified governance creates an organizational environment whereby the key stakeholders have a defined partnership with executive buy-in and oversight to create a uniform approach and to establish a strong linkage between legal obligations for information, records management, and IT; and the duty and value associated with the data asset.

Source: IGRM White Paper

Uniform Resource Identifier (URI)

A symbolic string representation of the location of an internet resource. A URL, Uniform Resource Locator is one type of URI for objects on the World Wide Web. "Web address" is a URL that uses the HTTP or HTTPS protocol.

Source: Herb Roitblat, Search 2020: The Glossary.

Unitization

The assembly of individually scanned pages into documents:

- **Physical unitization** utilizes actual objects such as staples, paper clips and folders to determine pages that belong together as documents for archival and retrieval purposes.
- **Logical unitization** is the process of human review of each individual page in an image collection using logical cues to determine pages that belong together as documents. Such cues can be consecutive page numbering, report titles, similar headers and footers and other logical cues.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Universal Serial Bus (USB)

A Plug and Play interface between a computer and a peripheral such as a mouse, keyboard, digital camera, printer or scanner. Unlike devices connected via SCSI ports, USB devices can be added to and removed from the computer without having to reboot the computer.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

UNIX

Pronounced yoo-niks, a popular multi-user, multitasking operating system developed at Bell Labs in the early 1970s. Created by just a handful of programmers, UNIX was designed to be a small, flexible system used exclusively by programmers.

Source: http://www.webopedia.com/TERM/U/UNIX.html

An operating system developed by Bell Laboratories that offers multi-user functionality and uses high-level programs. On PCs, it is often marketed under the name Xenix.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A software operating system. Originally pioneered by Bell Labs – now widely used by workstations.

Source:Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

DOS	Network operating system	Windows
Linux	NOS	Xenix
Microsoft DOS	Operating system	
Microsoft Windows	OS	

Unstructured Data

Data that is not in tabular or delimited format. File types include word processing files, html files (web pages), project plans, presentation files, spreadsheets, graphics, audio files, video files and emails.

Source: RenewData, Glossary (10/5/2005).

Unsupervised Learning

A Machine Learning method in which the learning Algorithm infers categories of similar Documents without any training by Subject Matter Expert(s). Examples of Unsupervised Learning methods include Clustering and Near-Duplicate Detection.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

A kind of machine learning where the objects are not labeled by an exterior source. Instead, the machine learning system organizes the objects based on implicit criteria that it derives. The selection of criteria is a function of the specific learning methods that are employed, the nature of the objects, and the way in which features of the object are represented. Clustering is an example of an unsupervised machine learning method. The goal of unsupervised learning is typically to identify hidden structure in unlabeled data, to summarize key features of the data.

Source: Herb Roitblat, Search 2020: The Glossary.

Upload

To transfer data from a user's computer to a remote computer system.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

URI

See: Uniform Resource Identifier (URI)

USB

See: Universal Serial Bus (USB)

USB Flash Drive

See: Jump Drive

User Created File

A file that was created because of the actions of the user, with or without intent or awareness. Excludes system generated log files.

Source: David Greetham, greethamdavid@yahoo.com (2008).

Data created by a person or a person's interaction with a computer.

Source: John Martin, johnmartin_va@comcast.net (2008).

User Group

Any organization made up of computer users (as opposed to vendors) designed to give the users a forum to share information about a particular system.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

User Guide

A set of instructions or a manual for a software program or hardware system.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

User-Friendly

Term used to describe a software program that is both easy to learn and easy to use.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

UTBMS: Conversion to Production Format

The process of restoring data that has been "deleted" from a storage device or retrieving data from a device that has failed, been corrupted, is damaged, or considered inaccessible.

Source: EDRM Metrics Glossary

UTBMS: Data Recovery

The process of restoring data that has been "deleted" from a storage device or retrieving data from a device that has failed, been corrupted, is damaged, or considered inaccessible.

Source: EDRM Metrics Glossary

UTBMS: Data Steward

A data steward is someone that is responsible for maintaining and managing the data assets of a particular organization. The role of data steward can be contrasted from a data custodian in that, though they both may share certain responsibilities with regards to data, a data custodian, in the e-discovery context, is often used to describe the individual responsible for the day-to-day control of a certain data set (i.e. an individual is the data custodian for their email inbox, an IT manager is the data custodian for file shares on a network server).

Source: EDRM Metrics Glossary

UTBMS: Defensibility

A highly prized Information Governance (IG) solution attribute today and in the foreseeable future is Defensibility. CE-discovery and archiving vendors tout this concept, especially as it relates to an entity's litigation or regulatory activities and its ability to produce, in a timely fashion, written documents or ESI. However, in practice, defensibility is a much broader concept.

In the parlance of the IG space, defensibility can apply to:

- The ability to demonstrate that appropriate, achievable, consistent policies governing the management of physical and electronic records have been developed and

implemented and that employees have been informed and educated on those policies as well as offered ongoing training and updates.

- The ability to demonstrate repeatable processes that support a firm's need to comply with legal or regulatory requirements.
- The ability to respond to legal or regulatory ediscovery requests in a timely fashion to thwart questionable litigations or potential fines that could be levied due to inability to produce ESI.
- The implementation of solutions that offer predictability whether it be to support compliance with retention policies, the ability to capture all appropriate ESI or the ability to scale as more content is managed electronically with assurance that all needed ESI is captured and preserved.
- The creation of an information security strategy that limits both external and internal risks and breaches when they occur.
- A risk management strategy that identifies potential liabilities, improves disaster preparedness and protects corporate and personal assets.

Source: *http://wikibon.org/wiki/v/Not_your_Fathers_Enterprise_Information_Archiving_Solution:_The_Next_Generation_Defined*

Source: EDRM Metrics Glossary

UTBMS: ESI Data Map

Data mapping finds or suggests associations between files within a large body of data, which may not be apparent using other techniques.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

UTBMS: ESI Inventory

An ESI inventory is a systematic process for identifying all of the records and non-record information in an organization, who creates, uses, or receives the information, and where users store it. A completed inventory provides a complete picture of the information environment. This picture is very helpful for assessing the needs of your RIM program.

Source: http://www.aiim.org/community/blogs/expert/carrying-out-a-records-inventory#sthash.TUKP9zzt.dpuf

UTBMS: ESI Preparation

Preparing ESI (electronically stored information) for processing and presentation.

Source: EDRM Metrics Glossary

UTBMS: ESI Presentation

Presenting ESI (electronically stored information) to the desired viewer, whether it be in depositions, production, trial, etc. ESI may be presented in native format, near native format, or in some other form acceptable to the parties.

Source: EDRM Metrics Glossary

UTBMS: ESI Processing

Any action taken on data using technology to reduce a data corpus based on specific criteria, organize data according to certain parameters, or convert data to another format more suitable for review and analysis.

Source: EDRM Metrics Glossary

UTBMS: ESI Staging

Data staging is the process by which original ESI files are copied, isolated, and stored in a forensically sound manner for future use.

Source: EDRM Metrics Glossary

UTBMS: Exception Handling

Exception handling is the process of responding to the occurrence, during computation, of exceptions – anomalous or exceptional events requiring special processing – often changing the normal flow of program execution. It is provided by specialized programming language constructs or computer hardware mechanisms. Exceptions also occur when doing a review, for example.

In general, an exception is handled (resolved) by saving the current state of execution in a predefined place and switching the execution to a specific subroutine known as an exception handler. If the exception state permits continuation, the handler may later resume the execution at the original state using the saved information. For example, a floating point divide by zero exception will typically, by default, allow the program to be resumed, while an out of memory condition might not be resolvable transparently.

Alternative approaches to exception handling in software are error checking, which maintains normal program flow with subsequent explicit checks for contingencies reported, using special return values or some auxiliary global variable such as C's errno or floating point status flags; or input validation to preemptively filter exceptional cases.

Source: http://en.wikipedia.org/wiki/Exception_handling. For more information on exception handling, see http://www.meridiandiscovery.com/articles/exception-handling-and-reporting-in-e-discovery/.

Source: EDRM Metrics Glossary

UTBMS: First Pass Document Review

Where a document review is organized in stages, the first pass document review is the first look at the documents that were identified as potentially responsive or relevant from the initial document collection. Typically, a first pass reviewer analyzes the documents for relevance or responsiveness and codes or marks them as such. Often, the reviewer will code for confidentiality and make an initial privilege determination during the first pass review.

Source: EDRM Metrics Glossary

UTBMS: Forensic Analysis Activity

Forensic analysis is the use of controlled and documented analytical and investigative techniques to identify, collect, examine, and preserve digital information. Recognizing the fragile nature of digital data, and the legal and regulatory requirements to properly preserve electronically stored information (ESI) during forensic investigations.

Source: EDRM Metrics Glossary

UTBMS: Hosting Costs

The cost to host data on a database or review platform; traditionally, the hosting phase occurs after data is collected, processed, and loaded to the review tool. Cost of hosting is typically by GB per month.

Source: EDRM Metrics Glossary

UTBMS: Legal Hold

A legal hold is a communication issued as a result of current or anticipated litigation, audit, government investigation or other such matter that suspends the normal disposition or processing of records. Legal holds can encompass business procedures affecting active data, including, but not limited to, backup tape recycling. The specific communication to business or IT organizations may also be called a “hold,” “preservation order,” “suspension order,” “freeze notice,” or “hold notice.”

Source: Sharon D. Nelson, Bruce Olson, John W. Simek, The Electronic Evidence and Discovery Handbook: Forms, Checklists and Guidelines, American Bar Association Law Practice Division (2006).

Source: EDRM Metrics Glossary

UTBMS: Native Format

Electronic documents have an associated file structure defined by the original creating application. This file structure is referred to as the “native format” of the document. Because viewing or searching documents in the native format may require the original application (i.e., viewing a Microsoft Word document may require the Microsoft Word application), documents are often converted to a standard file format (i.e., tiff) as part of electronic document processing.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Source: EDRM Metrics Glossary

UTBMS: Near-Line Storage

Near-line storage is used as an inexpensive, scalable way to store large volumes of data. Near-line storage devices can include DAT and DLT tapes, optical storage, and standard also slower P-ATA and SATA hard disk drives. Near-Line implies that the storage is not immediately available, but can be made online quickly without human intervention. Near-line can be slower, but generally, the type of data stored in near-line systems does not require instant access.

Source: http://www.webopedia.com/TERM/N/near-line_storage.html.

Source: EDRM Metrics Glossary

UTBMS: Near-Native Forms

A near-native format describes an electronic document that has been altered or converted from its original form in order to provide enhanced content control to a producing party while maintaining a level of usability consistent with its original format (e.g. conversion of a word document to a TIFF image with OCR to support redactions).

Source: EDRM Metrics Glossary

UTBMS: Non-Custodial Data

Data or records that are not created or maintained by an individual user, or whose physical storage and protection during the retention cycle are maintained by a system custodian and not end-users. Examples of non-custodial data may include data in certain structured systems, or access control or similar logs. It may not be possible to attribute authorship to non-custodial data. In contrast to Custodial Data.

Source: EDRM Metrics Glossary

UTBMS: Objective Coding

The recording of basic data such as date, author, or document type, from documents into a database.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Extracting information from electronic documents such as date created, author recipient, CC and linking each image to the information in pre-defined objective fields. In direct opposition to Subjective coding where legal interpretations of data in a document are linked to individual documents. Also called bibliographic coding.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Extracting various segments of information from a document such as its author, recipient, mailing date, or other fields, etc. Objective Coding is usually done from the document text or image because metadata or searchable text may be unavailable (e.g. a handwritten document that has been scanned), or the document may contain inaccurate metadata (e.g. metadata associated with a document written and signed by a partner might reflect the administrative assistant as the author where the document was originally typed on the assistant's computer).

See also:

Bibliographic Coding	Issue coding	Taxonomic coding
Coding	Level coding	Verbatim coding
Indexing	Subjective coding	
Issue Code	Tag	

UTBMS: Off-Line Storage

Any storage medium that is not immediately available, and must be inserted into a storage drive by a person before it can be accessed by the computer system. Examples include CD/DVD optical media, USB memory sticks, and tape cartridges. Offline storage is also called removable storage.

Source: http://www.webopedia.com/TERM/O/offline_storage.html.

Source: EDRM Metrics Glossary

UTBMS: On-Line Storage

Online storage is fully accessible and immediately available. This includes DRAM memory, solid-state drives (SSD), and always-on spinning disk, regardless of rotational speed. In contrast to near-line storage and off-line storage.

Source: EDRM Metrics Glossary

UTBMS: Preservation Order

Also called a "legal hold," "hold," "hold order," "hold notice," "suspension order," or "freeze notice". A Preservation Order is a communication issued as a result of pending or reasonably anticipated litigation or government investigation or action directing the suspension of the normal disposition or processing of records, including electronically stored information.

Source: EDRM Metrics Glossary

UTBMS: Privilege Review

A review of the documents identified as responsive or relevant in a particular legal proceeding for the additional legal classification of privilege whether as attorney-client communication or under the work-product doctrine. The law permits a disclosing party to withhold production of

documents on the grounds of legal privilege. Usually, a Privilege Log is generated in conjunction with the Privilege Review. See also Privilege Log.

Source: EDRM Metrics Glossary

UTBMS: Production Format

The format in which various documents are delivered from one party to another during the course of a legal proceeding. Available formats for document production are native, near or quasi-native, image (e.g. TIFF or PDF images), and paper. Rule 26(f) sets an expectation that the method and format by which ESI is to be produced should be considered and negotiated by the parties early in the discovery process. FRCP 34(b)(1)(E)(ii) states that "if a request does not specify a form for producing ESI, a party must produce it in a form or forms in which it is ordinarily maintained or in a reasonably usable form or forms." A key question regarding production formats is whether to include associated metadata. See also: native, near-native, image, and paper.

Source: EDRM Metrics Glossary

UTBMS: Project Management

Project management is the discipline of planning, organizing, motivating, and controlling resources to achieve specific goals.

Source: http://en.wikipedia.org/wiki/Project_management.

A project is a temporary endeavor with a defined beginning and end, undertaken to meet specific goals and objectives and can be distinguished from operations (business as usual). The primary objective of project management is to deliver the project goals while managing the constraints on project delivery. The primary constraints are scope, time, quality and budget.

Source: PMI (2010). A Guide to the Project Management Body of Knowledge p.27-35.

The secondary challenges are to optimize the allocation of necessary inputs and integrate them to meet pre-defined objectives.

Source: http://en.wikipedia.org/wiki/Project_management.

Project management principles apply to e-discovery, and most e-discovery activities are projects.

UTBMS: Redaction

The processing of editing the content of a document, usually by obscuring or removing certain sensitive, confidential or privileged information, prior to its production from one party to another.

Source: EDRM Metrics Glossary

UTBMS: Second Pass Document Review

Where a document review is organized in stages, the second-pass document review is the second, more detailed review of documents that were identified as potentially responsive or relevant in the first-pass review. Second-pass review can consist of a detailed review of documents to determine what documents should be withheld production on the grounds of privilege, relevance or other factors and which documents should be redacted. Second-pass review can also be used to quality-check (QC) the first-pass review. Second-pass review is frequently performed by more senior attorneys. In contrast to first-pass review. See also: Document Review.

Source: EDRM Metrics Glossary

UTBMS: Secondary Line Storage

Computer storage, as on disk or tape, supplemental to and slower than main storage, and not under the direct control of the CPU and generally contained outside it.

Source: http://dictionary.reference.com/browse/secondary+storage.

Source: EDRM Metrics Glossary

UTBMS: Structured Data

Structured Data is data that is organized. The most common type is database content. It refers to any type of data organized such as Internet data or other types of data that has been indexed.

Source: EDRM Metrics Glossary

Data that resides in a fixed field within a record or file is called structured data. This includes data contained in relational databases and spreadsheets.

Source: http://www.webopedia.com/TERM/S/structured_data.html.

UTBMS: Subjective Coding

The Subjective Coding of a document involves linking a legal interpretation to an individual document. In direct opposition to objective coding, in which bibliographic data about the document is recorded. Subjective Coding types include the classification of documents as privileged and responsive, and the categorization of documents by legal issue (“issue coding”).

Source: EDRM Metrics Glossary

Entering information from a document that requires the coder to exercise judgment, such as subject or issue codes. This field is often left blank for the law firm’s paralegals or associates to fill in.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

The coding of a document using legal interpretation as the data that fills a field. Performed by paralegals or other trained legal personnel.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Categorizing documents by their responsiveness to specific case issues or topics.

See also:

Bibliographic Coding	Issue coding	Taxonomic coding
Coding	Level coding	Verbatim coding
Indexing	Objective coding	
Issue Code	Tag	

UTBMS: Unstructured Data

Unstructured Data is the majority of data created today. It is the opposite of "structured data" such as indexed data found in a database because it is not pre-organized or pre-defined. Examples of unstructured data include Microsoft Word and other word processing documents; spreadsheets; email; Web pages; images; videos; and text.

Source: EDRM Metrics Glossary

Utilities

A set of routines designed to service a program or system. Examples are utilities file maintenance, information recovery from damaged disks, disk initializing, disk copying, routine system maintenance checks, and supervisory functions.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

V

V.32bis

The ITU standard for 14.4 kbs modem communications.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

V.34

The proposed ITU standard for 28.8 kbs modem communications.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

VAD

See: VAR

Validation

The act of confirming that a process has achieved its intended purpose. Validation may involve Statistical or Judgmental Sampling.

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

Validation Methodologies

Validation methodologies involve the case team in reviewing samples of documents to determine litigation relevance to classify documents as Responsive or Not Responsive to the issues of the case and therefore increasing the precision of the search results. Results of a keyword or iterative search may be validated by observing the frequency of hits, validating dropped items, sampling non-hits, and review call feedback analysis.

Source: EDRM Search Glossary.

Validation Table

Also called a "lookup table." A pre-defined set of entries for a specific field, often abbreviations, which appear when the coder moves to that field. Validation tables are used to cut down on errors during data entry.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Value

Utility or business purpose of specific information. The line of business has an interest in information proportional to its value — the degree to which it helps drive the "Profit" or purpose of the enterprise itself, its mission and goals.

Source: IGRM White Paper

Value-Added Dealer

See: VAR

Value-Added Reseller

See: VAR

Value-Added Specialty Distributor

See: VAR

VAR

Value-Added Reseller
Value-Added Dealer
Value-Added Specialty Distributor

Companies or people who sell computer hardware or software and "add-value" in the process. Most usually the value added is specific technical or marketing knowledge and/or experience.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

VASD

See: VAR

VDT (Video Display Terminal)

Generic name for all display terminals.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Vector Graphics

A set of routines designed to service a program or system. Examples are utilities file maintenance, information recovery from damaged disks, disk initializing, disk copying, routine system maintenance checks, and supervisory functions.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Vendor

The seller of computers or applications.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Vendor-Added Metadata

Data created and maintained by the electronic discovery vendor as a result of processing the document. While some vendor-added metadata has direct value to customers, much of it is used for process reporting, chain of custody, and data accountability. Contrast with customer-added metadata.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Customer-added metadata	Extrinsic data	File-specific metadata
Document metadata	File parameters	General metadata
Email metadata	File system metadata	Metadata

Verbatim Coding

Extracting data from documents in a collection in a way that matches exactly as the information appears in the documents. The opposite of the standardization type coding treatment.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Bibliographic Coding	Issue coding	Tag
Coding	Level coding	Taxonomic coding
Indexing	Objective coding	
Issue Code	Subjective coding	

Version

A release of a software program. Normally, new versions include additional features.

Vertical Deduplication

Deduplication within a custodian; identical copies of a Document held by different custodians are not Deduplicated. (*Cf*. Horizontal Deduplication.)

Source: Maura R. Grossman and Gordon V. Cormack, EDRM page & The Grossman-Cormack Glossary of Technology-Assisted Review, with Foreword by John M. Facciola, U.S. Magistrate Judge, 2013 Fed. Cts. L. Rev. 7 (January 2013).

See also:

Basic de-duplication	De-duplication	Global Deduplication
Case de-duplication	Duplicate	Horizontal Deduplication
Custodian de-duplication	Dynamic de-duplication	Production de-duplication

VESA (Video Electronics Standards Association)

Concentrates on computer video standards.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

VGA (Video Graphics Adapter)

A PC industry standard, first introduced by IBM in 1987, for color video displays. The minimum dot (pixel) display is 640 by 480 by 16 colors. Then "Super VGA" was introduced at 800 x 600 x 16, then 256 colors. VGA can extend to 1024 by 768 by 256 colors. Replaces EGA, an earlier standard and the even older CGA. Newer standard displays can range up to 1600 by 1280.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Video Display Terminal

See: VDT (Video Display Terminal)

Video Electronics Standards Association

See: VESA (Video Electronics Standards Association)

Video Graphics Adapter

See: VGA (Video Graphics Adapter)

Video Scanner Interface

A type of device used to connect scanners with computers. Scanners with this interface require a scanner control board designed by Kofax, Xionics or Dunord.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Videoblog (Vlog)

A vlog is a Weblog that uses video as its primary medium for distributing content. Vlog posts are usually accompanied by text, image, and other metadata to provide a context or overview for the video.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Virtual

The creation of case-specific words and codes to ensure uniform data entry. Used in conjunction with validation tables.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Virtual Private Network (VPN)

A way to provide remote access to an organization's network via the Internet. VPNs send data over the public Internet through secure "tunnels."

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A virtually private network that is constructed by using public wires to connect nodes.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Vlog

See: Videoblog (Vlog)

Vocabulary Control

The creation of case-specific words and codes to ensure uniform data entry. Used in conjunction with validation tables.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Volume

Volume is the principal variable of Metrics. Volume is the amount of data that is part of the eDiscovery collection. Volume will set the estimate for Cost and Time. For example, a large data Volume will cause an increase in the Time required to complete the Phases for Processing, Review and Production, thus increasing the Cost of the project. If the volume of data decreases, Time & Cost will also likely decrease.

Source: EDRM Metrics Glossary

VPN

See: Virtual Private Network (VPN)

W

WAIS (Wide Area Information Server)

A central database used for information access by network users in multiple physical locations. Often refers to an Internet database, but WAIS servers have existed for some time outside the Internet arena.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

Database	Full text database	SQL
Flat file database	Relational database	

WAN (Wide Area Network)

A central database used for information access by network users in multiple physical locations. Often refers to an Internet database, but WAIS servers have existed for some time outside the Internet arena.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A system of local area networks in different physical locations connected through communications software.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Generally a network of PC's, remote to each other, connected by telecommunications lines.

See also:

Client/server network	Network	Stand alone computer
LAN - local area network	Peer-to-peer network	
MAN - metropolitan area network	SAN - storage area network	

WAP (Wireless Application Protocol)

A widely used set of protocols that standardize the manner in which wireless devices, such as cell phones and some PDAs are able to access parts of the Internet, such as e-mail and the Web.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

WAV

File extension name for Windows sound files. Compression is not required. .WAV files can reach 5 megabytes for one minute of audio.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Web (World Wide Web)

The WWW is made up of all of the computers on the Internet which use HTML-capable software (Netscape, Explorer, etc.) to exchange data. Data exchange on the WWW is characterized by easy-to-use graphical interfaces, hypertext links, images, and sound. Today the WWW has become synonymous with the Internet, although technically it is really just one component.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

The portion of the Internet with a GUI.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Web Review

Allows clients and litigators to review metadata, text, images, or native files, in any combination, via a Web browser connected to a document repository via a secure on-line connection.

Source: Ibis Consulting, Glossary.

Web Site

A collection of Uniform Resource Indicators (URIs, including URLs (Uniform Resource Locators)) in the control of one administrative entity. May include different types of URIs (i.e., file transfer protocol sites, telnet sites, as well as World Wide Web sites).

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

Weighted Relevance Search

A type of search that will allow the user to sort and retrieve documents according to a statistical "weight" given by the use of a mathematical relevancy evaluation program.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Wildcard search

What You See Is What You Get (WYSIWYG)

Pronounced "wizeewig." A system that allows the user to see on screen exactly what will be printed out.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Display & software technology which shows on the computer screen exactly what you'll get when you print that screen. Usually requires a large, high-density monitor.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Wide Area Information Server

See: WAIS (Wide Area Information Server)

Wide Area Network

See: WAN (Wide Area Network)

Wildcard Search

The wildcard symbol, typically "*", can be used with any other search to retrieve different variations of the same word, e.g., “insur*” for insurance, or insured.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Ad Hoc Search
Adaptive pattern recognition
Associative retrieval
Boolean search
Combined word search
Compliance Search
Concept search
Exploratory Search
Full text search
Fuzzy search
Index
Index/coding field
Keyword
Keyword search
Natural language search
Numeric range search
Phonic search
Phrase search
Proximity search
Range search
Search
Similar document search
Sound-alike
Stemming
Synonym search
Term search
Topical search
Weighted relevance search

Wildcards

Symbols such as * or ? included within a Keyword to indicate that the location where the symbols are used may match a single letter or multiple letters.

Source: EDRM Search Guide Glossary.

Source: EDRM Search Glossary.

Windows

A software product that provides an operating environment that runs under MS-DOS, using a GUI that can run different programs at the same time in different windows.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

DOS	Network operating system	UNIX
Linux	NOS	Xenix
Microsoft DOS	Operating system	
Microsoft Windows	OS	

Windows NT File System

See: NT Filing System (NTFS)

WinZip

WinZip is a program commonly utilized to zip files.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Wipe

Term for deliberately overwriting a piece of media and removing any tract of files or file fragments.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Wireless Application Protocol

See: WAP (Wireless Application Protocol)

Wordnet

An electronic thesaurus developed by George Miller and his students at Princeton University. Used by some systems to provide synonyms for query expansion.

Source: Herb Roitblat, Search 2020: The Glossary.

Workflow

The stream of information processing through an organization.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Workgroup

A group of computer users connected to share individual talents and resources as well as computer hardware and software – often to accomplish a team goal.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Workstation

A single computer, either a desktop with a hard disk or a dumb terminal.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

A powerful microcomputer or minicomputer with a RISC chip, typically used by engineers or graphics technicians.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

See also:

Computer	Microcomputer	Personal computer
File server	Minicomputer	
Laptop computer	Notebook computer	

World Wide Web

See: Web (World Wide Web)

WORM (Write Once, Read Many)

An optical disc storage device that uses laser technology similar to the CD-ROM. Information written to the WORM disc, cannot be altered. The advantages of WORM are increased disc density and life expectancy.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

Data storage devices (e.g. CD-ROM's) where the space on the disks can only be written once. The data is permanently stored. This is often today's primary media for archival information. The expected viable lifetime of a WORM is at least 50 years. Since it's impossible to change, the government treats it just like paper or microfilm and it is accepted in litigation and other record-keeping application. On the negative side, there is no current standard for how WORM's are written. The only ISO standard is for the 14" version, manufactured only by one vendor. A 5.25" standard is emerging from the European Computer Manufacturing Association but is not yet accepted. Further, WORM discs are written on both sides, but there are currently no drives that read both sides at the same time. As for speed, WORM is faster than tape or CD-ROM, but slower than magnetic. Typical disk access times run between 40 and 150 milliseconds (compared with 11 ms for fast magnetic disks and 300 ms for CD-ROM. Data transfer rates run between 1 and 2 MB/sec (compared with 5 to 10 for magnetic discs and 600KB/sec for CD-ROM. Disk sizes run from 5.25" (1.3 gigabytes) to 12" (8 to 10 gigabytes) capacities. There is

also a 14'" disc (13 to 15 gigabytes), only manufactured by Kodak's optical storage group. WORM's can also be configured into jukeboxes. There are various technologies:

Technology Description	*Benefit*	*Drawback*
Ablative: Laser burns holes in disk	Unalterable data	Dust, moisture may affect media
Bubble-forming: Laser forms bubbles in the media	Unalterable data	Few drives available
Dye Polymer: Laser heats dyed layer to form bumps	Potential low media cost	Laser mechanism more expensive; disks wear out faster; few drives available
Magneto: Laser focuses magnetic field	Many suppliers, long disk life	No true WORM in multi-function; data theoretically alterable
Phase change: Laser heat changes disk's molecular structure	One-pass data (no erase step)	Same as Dye Polymer

From Imaging Magazine, September, 1994

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

WORM Disk

A popular archival storage media during the 1980s. Acknowledged as the first optical disks, they are primarily used to store archives of data that cannot be altered. WORM disks are created by standalone PCs and cannot be used on the network, unlike CD-Rs.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

See also:

CD
CD-R
CD-ROM
CD-RW
Disc
Disk
Diskette
DVD
DVD-ROM
Floppy disk
Hard disk
Hard drive
Jaz disk
Laser disc
Magnetic disk
Magnetic storage media
Media
Optical disk
Storage media
Zip disk

Write Once, Read Many

See: WORM (Write Once, Read Many)

Write Protect

Restrict a diskette from having information recorded to it. Used to prevent the erasure of valuable information.

Source: Legal Electronic Document Institute, Basic Principles of Automated Litigation Support (2005).

WYSIWYG

See: What You See Is What You Get (WYSIWYG)

X

X.25

A standard protocol for data communications.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

Xerography

A beam of light hits an electrically charged drum and causes a discharge at that point. Toner is then applied which sticks to the non-charged areas. Paper is pressed against the drum to form the image and is then heated to dry the toner. Used in laser printers and copying machines.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

XML

See: Extensible Markup Language (XML)

Y

Yield

See Prevalence or Richness.

Source: The Grossman-Cormack Glossary of Technology Assisted Review (Version 1.02, Nov. 2102).

Z

Z-Test of Proportions

A statistical hypothesis test comparing two proportions. This test assumes that the two proportions are distributed approximately according to the normal distribution. With sample sizes more than a few tens, this is an appropriate assumption. It tests the hypothesis that there is no difference between the two population proportions.

Source: Herb Roitblat, Predictive Coding Glossary.

Zero-Length File

A file with file properties, 0 byte size, (0B or 0K) but no content, or metadata (including commercial application data).

Source: Ibis Consulting, Glossary.

Zip

The act of compressing large files into a single file, called a zip file. Zip files take up less storage space so they easy to send via email. Not to be confused with Zip drive, a portable storage peripheral.

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

A common file compression format that allows quick and easy storage for transport. Compresses and combines one or more documents by utilizing an algorithm that 'removes' white space and replaces it when decompression takes place. Commonly used to combine and send large documents via email.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

An open standard for compression and decompression used widely for PC download archives. ZIP is used on Windows-based programs such as WinZip and Drag and Zip. The file extension given to ZIP files is .zip.

Source: Kroll Ontrack, Glossary of Terms, http://www.krollontrack.com/glossaryterms

An algorithm used to create a compressed archive. The archive can contain many different files, each of which can be recovered by “unzipping” the archive.

Zip Disk

A large capacity floppy disk that can only be read from or written to using a proprietary zip disk drive.

See also:

CD
CD-R
CD-ROM
CD-RW
Disc
Disk
Diskette
DVD
DVD-ROM
Floppy disk
Hard disk
Hard drive
Jaz disk
Laser disc
Magnetic disk
Magnetic storage media
Media
Optical disk
Storage media
WORM disk

Zip Drive

A brand-name magnetic storage device that can hold between 100 and 250 megabytes of data.

Source: Fios, E-Discovery Glossary, http://discoveryresources.org/01_electronic_discovery_glossary.html

Source: Vinson & Elkins LLP Practice Support, EDD Glossary.

Source: RSI, Glossary.

See also:

Disk drive	Magneto-optical drive	Tape drive
Floppy disk drive	Portable drive	
Jaz drive	Storage device	

Zone OCR

An add-on feature of the imaging software that populates document templates by reading certain regions or zones of a document, and then placing the text into a document index.

Source: Formerly American Document Management, Glossary of Terms, now 5i Solutions Glossary.

.E01 File

".E01" is a legacy EnCase evidence file format. An ".E01" file is a byte-for-byte representation of a physical device or a logical volume.

Source: EnCase Forensic Imager, Version 7.06, User's Guide. Guidance Software.

.Ex01 File

".Ex01" is the current EnCase evidence file format. An ".Ex01" file is a byte-for-byte representation of a physical device or a logical volume. It has LZ compression, AES256 encryption with keypairs or passwords, and options for MD5 hashing, SHA-1 hashing, or both.

Source: EnCase Forensic Imager, Version 7.06, User's Guide. Guidance Software.