UNLOCK THE VALUE OF YOUR DATA USING eDISCOVERY TECHNOLOGY IN ARBITRATIONS

Amit Jaju & Ankush Lamba[*]

I. BACKGROUND

In today's digital era, organizations are generating an unprecedented volume of communications data and documents in varied forms spread across servers, cloud, laptops, mobile phones, etc. According to industry reports, 79 zettabytes of data will be created in 2021 and this figure is projected to grow to over 180 zettabytes by 2025.[1]

"***Data is the new oil.*** Data is valuable but just like oil, it is not the raw data that's worthwhile, but the refined data obtained by breaking it down and analysing thereafter."[2] When properly organised and refined, it provides invaluable insights.

In case of a dispute, it is now common that the relevant data exists mostly in electronic form, which underscores the significance of electronic data being the primary gateway to facts surrounding the case. The parties, thus, preparing for an arbitration should be aware of the rules governing the use of electronically stored information ["**ESI**"] and explore electronic discovery ["**eDiscovery**"] technology to provide fact-based assessments on large volumes of data in a timely and effective manner.

eDiscovery refers to discovery in legal proceedings such as arbitration, litigation, government investigations, or Freedom of Information Act requests, where parties involved preserve, collect, process, review, analyse and report information in electronic formats for the purpose of using it as evidence. All these aforementioned steps are facilitated by eDiscovery technology and processes; enabling the parties to unveil important information concerning the dispute with reduced costs, quickened decisions, and mitigated risks.

---

[*] Amit Jaju is a Senior Managing Director in Ankura Consulting Group, LLC and is the country leader for India, and Ankush Lamba is a Managing Director with Ankura Consulting Group, LLC in the Data & Technology practice.

[1] Arne Holst, *Volume of Data/Information Created, Captured, Copied, And Consumed Worldwide From 2010 to 2025*, STATISTA (Jun. 7, 2021), https://www.statista.com/statistics/871513/worldwide-data-created/.

[2] Charles Arthur, *Tech Giants May Be Huge, But Nothing Matches Big Data*, THE GUARDIAN (Aug. 2013), https://www.theguardian.com/technology/2013/aug/23/tech-giants-data.

The focus of this paper is threefold:

- to cover protocols for the disclosure of the ESI in an arbitration matter;

- to outline ESI and stages of eDiscovery technology; and

- to illustrate the ways to use eDiscovery in arbitration matters.

II. ESI DISCLOSURE IN ARBITRATIONS

Arbitration is an alternative form of dispute resolution enabling the parties to reach an amicable resolution outside the courts. An arbitrator has the authority to demand electronic data to be presented in such cases. With ESI becoming increasingly prevalent in arbitrations[3], it has become vital for the arbitrators seeking large volumes of electronic disclosure ["**e-disclosure**"] to familiarize themselves with the protocols laid down by arbitral institutions like the Indian Council of Arbitration, International Institute for Conflict Prevention and Resolution ["**IICPR**"],[4] Chartered Institute of Arbitrators ["**CIArb**"], ICC, LCIA guiding on disclosure of the data in an electronic format.

These protocols in general cover two purposes. First, to assist arbitrators in addressing document disclosure by setting out general principles for dealing with requests regarding the disclosure of documents and electronic information. Second, to allow parties either when drafting an arbitral agreement, or after a dispute arises, to elect "certain modes of dealing with the disclosure of documents."[5]

As per the Protocol of the CIArb, some of the early considerations that arbitrators and parties must include for the e-disclosure process are:[6]

(i)   whether disclosure of electronic documents and information is likely to be requested

---

[3]*Using Technology and e-Disclosure*, GLOBAL ARBITRATION REVIEW (Sept. 3, 2021), https://globalarbitrationreview.com/guide/the-guide-evidence-in-international-arbitration/1st-edition/article/using-technology-and-e-disclosure#footnote-069.

[4] Homepage, INTERNATIONAL INSTITUTE FOR CONFLICT PREVENTION & RESOLUTION, https://www.cpradr.org/.

[5]*Protocol for E-Disclosure in International Arbitration*, CIARB INTL. ARB. PROTOCOL, https://www.ciarb.org/media/1272/e-iscolusureinarbitration.pdf.

[6] *CPR Protocol on Disclosure of Documents and Presentation of Witnesses in Commercial Arbitration*, IICPR, https://www.cpradr.org/resource-center/protocols-guidelines/protocol-on-disclosure-of-documents-presentation-of-witnesses-in-commercial-arbitration/_res/id=Attachments/index=0/CPR-Protocol-on-Disclosure-of-Documents-and-Witnesses.pdf; JOHN RANGE & JONATHAN WILAN, INSIDE THE MINDS: EDISCOVERY IN ARBITRATION, (Thomson Reuters/Aspatore, 2010).

by either party;

(ii)   the types of electronic documents in each party's control and the identity of the computer systems, electronic devices, storage systems and other media on which ESI is retained;

(iii)   steps that should be taken to preserve electronic information;

(iv)   Particular rules that govern disclosure and use of the ESI, such as the IBA Rules on the Taking of Evidence in International Commercial Arbitration;

(v)   whether agreements limiting the scope or extent of e-disclosure are desirable;

(vi)   Tools and techniques that would be useful to focus the electronic search and reduce its cost;

(vii)   whether special arrangements to ensure data privacy or to protect privilege (i.e., clawback provisions) are appropriate; and

(viii) any professional guidance that is necessary to assist the parties or tribunal with IT issues related to disclosure.

Further, Schedule 2 of the IICPR protocol highlights the "Modes of Disclosure of Electronic Information" and the four "modes" of disclosure for ESI.[7]

- "Mode A" provides minimal e-disclosure, limited to documents a party will rely on in support of its case, produced in paper or other reasonably usable forms.

- "Mode B" provides for disclosure, "in reasonably usable form," by a limited number of "designated custodians" (the actual number to be selected by the parties and/or the tribunal), covering ESI created between the date of signing of the agreement in dispute and the date of the request for arbitration, to be provided only from "primary storage facilities" having "reasonably accessible active data" (i.e., not from backup servers or tapes, PDAs, or voicemails).

- "Mode C" is the same as Mode B, but covers a larger number of custodians, a wider time period, and allows the parties to agree upon a showing of "special need and

---

[7] Homepage, EDRM, https://edrm.net/.

relevance disclosure of deleted, fragmented or other information difficult to obtain other than through forensic means."

- "Mode D" authorizes disclosure essentially similar to the U.S. Federal Rules of Civil Procedure, where all non-privileged electronic information "relevant to any party's claim or defence" is produced, subject only to limitations on "reasonableness, duplicativeness, and undue burden."

Parties selecting Modes B, C, or D must "meet and confer, before an initial scheduling conference with the tribunal, concerning the specific modalities and timetable for electronic information disclosure."

III. ESI AND STAGES OF eDISCOVERY TECHNOLOGY

With technology evolving at a rapid pace, ESI can exist in various electronic devices in a wide range of formats and types spread across an organization's databases and other data sources. It can include "emails, documents, chats, message, databases, voicemail and audio/video files, writings, drawings, graphs, charts, photographs, sound recordings, images, and other data or data compilations." It also includes any type of data from social media, instant messaging platforms or smartphone apps. That means that every post on Facebook, Twitter or a message sent over WhatsApp, Signal or Telegram can be considered as valuable information.

Another important part of electronic data is the metadata embedded in electronic files. Metadata is the information generated within a piece of electronic data that is 'data about data'. It exists within every digital file stored on physical electronic devices, such as your computer and smartphone, IOT device etc. The information contained within metadata includes author, creation date, history, software used to create it, usage of the device, digital footprint data, etc. Metadata often entails the story about the electronic data file and, therefore, is often a key focus of eDiscovery[8] and must be carefully preserved to prevent data spoliation. Even moving a file from one location to another location can modify the metadata information and question the authenticity of the ESI.

Significant efforts are being put to create workflows and establish best practice recommendations to help organisations manage ESI and meet the legal requirements. In 2005,

---

[8] eDiscovery Homepage, CYFOR, https://cyfor.co.uk/ediscovery/.

EDRM ["**Electronic Discovery Reference Model**"] was created, which outlines the stages of the eDiscovery process.

EDRM is designed to guide on gathering and assimilating of electronic data during the legal process. The EDRM framework is a conceptual standard for the eDiscovery process. This means that those following EDRM could engage some, but not all, of the steps outlined in the model and still successfully discover relevant data.

In recent years, the EDRM has grown to embrace the information governance lifecycle, which is now listed as the first stage of the EDRM model.

The EDRM helps to determine what data to preserve and how to manage the preserved data. It includes the following steps:[9]

- **Information Governance** – Setting policies, procedures, processes and controls implemented to manage and secure electronic information consistently at an enterprise level, supporting an organisation's immediate and future regulatory, legal and risk requirements.

- **Identification** – Locating potential sources of ESI & determining its scope, breadth & depth.

- **Preservation** – Ensuring that ESI is protected against inappropriate alteration or destruction.

- **Collection** – Gathering ESI for further use in the e-discovery process (processing, review, etc.).

- **Processing** – Reducing the volume of ESI and converting it, if necessary, to forms more suitable for review & analysis.

- **Review** – Evaluating ESI for relevance & privilege.

- **Analysis** – Evaluating ESI for content & context, including key patterns, topics, people & discussion.

- **Production** – Delivering ESI to others in appropriate forms & using appropriate

---

[9] *EDRM Model*, EDRM, https://edrm.net/resources/frameworks-and-standards/edrm-model/.

delivery mechanisms.

- **Presentation** – Displaying ESI before audiences (at depositions, hearings, trials, etc.), especially in native & near-native forms, to elicit further information, validate existing facts or positions, or persuade an audience.

IV. WAYS TO USE eDISCOVERY TECHNOLOGY IN ARBITRATION

eDiscovery technology is an apt answer in providing time and cost-effective solutions to arbitration cases. Following are the ways in which eDiscovery technology can be utilised in an arbitration.

A. Data Preservation and Managing Legal Holds

Data preservation is the foundation of the eDiscovery process and one cannot engage in discovery if the relevant data information is not well preserved. It involves steps such as identifying, securing and maintaining data as per arbitration requirements. Further, in the data preservation process, there is a heightened focus on maintaining ESI in an original and unaltered condition. Unlike hard copy documents, electronic documents also contain metadata. Since metadata can play a significant role in such cases, parties involved in an arbitration demand that organizations preserve it to ensure the authenticity of the data.

In case the data is not preserved with relevant metadata, the opposing party may raise objections and the data may not be admissible. Further, any event of 'Data Spoliation' - missing information, destruction or alteration of relevant data or the loss of discoverable information can be subject to severe penalties by the arbitration authority or the individual arbitrator.

To avoid 'Data Spoilation', parties involved in the matter implement and manage legal hold to initiate the data preservation activity. Legal hold is the practice and process of ensuring ESI is preserved during a legal matter. Parties with their technology experts map out the relevant devices that store data and further map it out to the list of individuals who may be holding these devices or data. These individuals are referred to as 'custodians' and are sent a legal hold notice to exempt the electronic data in question from any activity of deletion or alternation till the time a resolution is reached in the underlying matter.

Data preservation or collection of the ESI is neither a one size fit all endeavour, nor is it a process that is supported by a single technology. Different approaches exist for ESI collection

which require high end digital forensic tools and forensic technology expertise. The approach taken depends on the type of digital device from which data needs to be collected. For example, the procedure for acquiring electronic data from a computer hard drive is different from the procedure required to obtain electronic data from mobile devices, such as smartphones. Digital forensics collection procedure essentially involves the following steps:

- Step 1 – Seizure of the storage media.

- Step 2 – Acquiring the storage media forensically; that is, creating a forensic image of the media using forensic tools. This will create a bit-by-bit exact copy of the original data and ensure the original data is not altered.

- Step 3 – Analysing the forensic image and original data by calculating the Hash value using an algorithm. A matched hash value[10] ensures that a bit-by-bit mirror image is created of the original media.

- Step 4 – Another backup copy of the forensic image is created for preserving the evidence and further analysis.

Data preservation activity is traditionally done on-site and in person. However, the impact of COVID-19 and restrictions on travel has resulted in adopting a remote approach for ESI collection.

B. Data Processing and Searching Techniques for Relevant Information

With the data collected in the data preservation activity, identification of the pertinent ESI is one of the most challenging phases as it requires determining what information is preserved and identifying the most efficient way to sort through the potentially relevant information. Two techniques are employed on the data - 1) Processing and 2) Searching techniques. In addition to these two techniques, advanced analytics and technology assisted review has also been deployed in recent times. These advanced techniques are discussed in this article.

i. Data Processing

---

[10] *Hash Values are the DNA of Digital Evidence*, FORENSIC DISCOVERY (Aug. 2021), https://forensicdiscovery.expert/hash-values-are-the-dna-of-digital-evidence/. Hash Value is a function that calculates numeric value of a fixed length to uniquely identifies data.

Before any collected data can be searched or reviewed, it must first be "processed." With the range of potential ESI sources continually multiplying and creating different formats of electronic data, it is important to choose the right solution platform that can cater to all the data types, sizes, and structures of any language and convert the collected ESI into a simple data form that can later facilitate efficient review. Choosing the right platform and having the right processing specification in the processing stage is very important to get a useful result and eliminate the risk of missing out on important information.

During the data processing phase, data is injected into the platform where the following activities take place:

- Native files are unpacked/expanded. For instance, an email with the attachment becomes a separate record with a link established as parent email and child attachment.

- Metadata for each data file is extracted and normalised in a single form for further search analysis.

- Unique control ID number is assigned to each data item for identification by the platform.

- Optical Character Recognition ["**OCR**"] technique involving detection of text/characters from image files, scanned documents or PDFs makes the documents searchable by keywords.

- Processing time zone – Files processed of different time zone are converted to a single time zone format as per the setting specified during processing.

- Data Culling – Using the extracted metadata fields, documents are filtered based on the date and relevant data is filtered out as per the decided period of scope. This is termed as 'Data filtration'.

- Deduplication – Duplicate files are identified across multiple data sources/custodians. Further 'Dupe Custodian' field is created which captures the custodian name whose version of a file was deduped out, so it can be determined who else had a copy of the documents.

- deNISTing – removing system files, program files, and other non-user created data

from the ESI that do not contain any user-generated content and is not relevant to the discovery matter.

Once the processing steps are completed, the data can now be searched for relevant information.

ii. Searching

The searching technique is another critical step that helps to reduce the volume of data by identifying and applying an effective set of keywords on the data to produce a good, effective and responsive result set. These keywords or "search terms" are discussed and agreed upon between the parties and their attorneys. Since reviewing every item of electronic data is impractical, the data set needs to be reduced to a manageable size for the parties to identify the relevant data to analyse. Keyword searching techniques enable to narrow down metadata, documents and e-mails content, thereby, reducing large data sets down to only those items that contain the information that is relevant for the matter and segregating non-responsive documents.

The choice of terms is crucial for a successful result. Also, it is recommended to take guidance from eDiscovery technology experts in framing the terms that are acceptable to the solution platform.

Another method of searching data is 'concept search or concept clustering'. This technique focuses on related concepts within the document or email that is conceptually similar to the search phrase. This technique takes a group of data and then breaks it into groups of "similar documents". It is effective when applied during the review phase as the grouping of the documents is an effective way to navigate a responsive dataset. However, careful consideration must be given to how concept searching is applied as it does not confirm the presence or absence of the target data for which concept search is applied. Further, when information is presented in conceptual clusters, the attorney can review that information at a faster rate, saving both time and money.

For matters with a compressed time frame, the attorney can sometimes become overburdened by the dependency on keyword searches, mainly in cases where the keywords set out are providing false hits or missing out on extremely important documents or unnecessary review time. The search output can turn out to be effective by working closely with the eDiscovery

technology team and underlining the objective of the matter. This helps in formulating an effective approach to search the data and identify which advanced search features may suit the attorney's requirement.

C. Technology-Assisted Review

Over the years, organisations have been overwhelmed by the volume of data that they continue to create and receive. Zettabytes of digital data is stored on the server, cloud storage and personal devices. This era of "Big Data" has challenged organisations to better manage their ESI. A complex arbitration matter may encounter a scenario of examining through terabytes of the data involving countless emails and documents to understand the case and also reviewing the large production received from the other side. This may make it increasingly difficult for the parties to manage the cost of arbitration matters. Attorneys and parties further look forward to eDiscovery technology - 'Technology Assisted Review' ["**TAR**"] to aid in culling down the huge amounts of data to make the review process more practical and cost-effective.

TAR is the process of using machine learning algorithms to categorize each document as responsive or not, or to prioritize the documents from most to least likely to be responsive, based on a review performed by the legal professionals by coding of a small subset of the documents. TAR is faster and more accurate than manual review in which reviewers go through every document one by one (often referred to as "linear review"). TAR not only helps to prioritize a review but it may also be possible to exclude a large percentage of the emails and documents as unreviewed. This significantly reduces the time required for the manual review and ensures that no key document is missed. To use TAR, the machine learning model needs to be trained to achieve quality output. For this, the eDiscovery technology professional works closely with legal professionals and subject matter experts. Further, the technology professional plays a vital role in setting up the required framework for TAR and guiding the reviewer team on the necessary steps that need to be followed such as:

- **Setting up the protocol and educating the reviewer:** This involves building the reviewer coding rules that are taken into account for the use of TAR. For instance, to decide how to treat the coding of family documents during the TAR training process. Further, these rules are highlighted to the reviewers before starting the TAR review.

- **Coding documents:** Reviewers apply subjective coding decisions to documents to adequately train the TAR model e.g., Relevancy, Responsive or Not Responsive.

- **Predicting result:** This step involves applying information "learned" from the previous step of coding documents and classifying a large number of documents with pre-determined labels.

- **Testing and evaluating results:** This involves validating the results generated using the TAR model, in an effort to create a meaningful metric of TAR performance and determining if the TAR system has achieved the goals anticipated by the review team.

<u>D. Redacting Privileged or Sensitive Information</u>

Protecting both privileged information as well as personally identifiable information ["**PII**"] is of utmost importance in any arbitration matter and needs to be carefully and methodically reviewed and safeguarded. Before producing documents in paper or electronic form to the opponent party, it is important to ensure that any privileged or sensitive information is fully removed. In this regard, redaction is an important activity in the eDiscovery process. In the early days, a typical method of redaction was to do it manually by printing the documents, using a black marker to mask the information, then photocopy the marked-up pages several times to ensure complete obscuration before re-scanning them back into the system. As the volume of ESI has increased, this method of redaction has become cumbersome, expensive and even unrealistic given the disclosure deadlines. Another challenge the attorney's face today is applying redaction manually on the electronic data using tools or platforms that do not have the feature of creating a redaction log report. Without the redaction report, this could turn to be a tedious activity and can take hours if it is to be applied for a large number of documents.

eDiscovery technology has developed and implemented redaction tools to search for privileged or sensitive or private information and ensure that it is eliminated entirely - not just from the document but also from the metadata of the document. Using this technology saves a significant amount of time and expenditure, enabling parties to produce thousands of redacted documents with just a few clicks in the platform, in a few minutes. It keeps a track of the documents redacted during the review and creates a consolidated report at the end of the review providing details such as a list of redacted documents with their metadata and reason for the redaction submitted.

E. Advanced Analytics

With the rise in eDiscovery technology, attorneys are moving away from the standard linear review of custodians and documents (i.e., reviewing one document after another, ordered by date or keyword relevance) and adopting the analytics approach in the document review to bring more efficiency and accuracy to the eDiscovery process.

Advanced Analytics in the eDiscovery process can relate information about the data, provide greater insight for the document review, locate key documents more quickly, which can later result in key informed decision making by attorneys and help them in estimating the scope of the eDiscovery project.

eDiscovery analytics is highly advanced, constantly evolving and requires training. In collaboration with the trained eDiscovery technology experts, these tools can save significant time of the attorneys when added to the workflow of the project and achieve quality output at a faster rate. Some of the advanced analytics features include:

- **Near Duplicate Detection:** It involves the comparison of every document against each other to determine whether their similarity is greater than the set threshold. If greater than a threshold then the documents are identified as near duplicate.

- **Communication Analytics:** Provides a visual representation of the communications over a specific set of email documents. Shows patterns of conversations such as – Who's having the conversations, how often, what are they talking about, and how invested are they?

- **Email Threading:** It identifies email relationships - threads, people involved in a conversation, attachments, and duplicate emails - and groups them together so you can view them as one coherent conversation.

- **Audio and Video Discovery:** Adding transcription of the text to the metadata of the files to ease the analysis and review.

- **Personal data (PII) detection:** performs pattern identification analytics highlighting social security numbers, phone numbers, health records, etc., for redaction.

- **Image Recognition and Classification:** Uses machine learning algorithms to identify

and apply labels to images to facilitate search and filter without looking at every image.

VI. CONCLUSION

eDiscovery in an arbitration requires the parties to gather the necessary ESI for an arbitrator to work with the parties to reach useful resolutions. In such matters, it can be beneficial for the parties to collaborate with eDiscovery technology experts throughout the course of the arbitration process to manage all technical activities related to the electronic data. This also lowers the burden on the attorneys to get the optimum output in a shorter timeframe from all the necessary steps that need to be performed across the matter such as identification of the data sources, managing the legal hold and preserving the data.

eDiscovery technology when properly utilized with technology experts can reduce organizational risk while also providing substantial cost reduction.