Processing & Reducing

1 INTRODUCTION

1.1 The stage of processing and reducing your pool of documents is key. Like many of the other stages of e-disclosure (for example, collection, review and analysis) it is a delicate balancing act. On the one hand, it is essential to try to reduce the volume of documents which have to be read by the lawyers, as this is the most costly part of the process. However, on the other hand, it is very important to avoid accidently excluding a category of documents which may later prove to be relevant and/or responsive.

2 PROCESSING AND REDUCING THE POOL OF DOCUMENTS[1]

2.1 The main thing to keep in mind is that every case is different: there is no "one size fits all" approach. Therefore, before making any decisions on how to process and reduce your data, make sure that you have considered the following:

(a) Who are your key custodians (i.e. the key individuals involved in the matter) – you may have collected data from additional custodians who turn out not to be as important as you originally thought;

(b) When were your key custodians involved – this may affect the date ranges of the data to be considered. Also consider what other periods are relevant;

(c) What type of data have you collected – one of the options open to you at this stage is to remove unwanted file types. These may include files with "strange" file extensions which appear to be either corrupt or not supported by your system (for example, logos). However, be cautious - just because you discover that you have collected a category of file types with a "strange" file extension, does not mean that you should automatically discard it. Some specialised file types (such as, for example, those recorded by custom-made software on a power plant) may not be recognised by your IT team who are not specialists in this field. Go back to your client and make enquiries as to what these files may be;

(d) Is it worth carrying out objective coding i.e. extracting such information from a document as its author, its mailing date, etc. Objective coding is usually done from the document text or image, because the metadata may be inaccurate. For example, a document written and signed by a partner might show the administrative assistant as the author in the metadata, because it was originally typed on the assistant's computer.

(e) Make sure that you are on top of the key issues – these will affect the decisions that you make in relation to how you can reduce your pool of documents (see "keyword filtering" below).

---

[1] *The worked example of processing and reducing uses screenshots from a number of tools including, eDiscovery Tools, Clearwell, IConnect and Ringtail.*

(f)   What foreign languages are likely to be present in your documents – this will affect how you filter your data (for example, key words in English will not work on a data set where a large proportion of the documents is in German).

(g)   An important factor to also keep in mind is considering how much time you have available. Often, this will dictate the volume of your pool of documents.

2.2   Some of the ways to reduce the volume of your pool of documents include:

2.2.1   **De-duplication -** removing duplicates from the data to be reviewed is essential. Technology exists which can identify exact duplicates so that these can be removed for the review stage but added back in for the Listing stage. This is known as de-duplication. Technology now also exists which can identify and strip out duplicates which emanate from different time zones (by converting all times to GMT time). Technology also exists which can find "near duplicates". *Example provided by Kroll.*

2.2.2   **Keyword filtering** - you should give consideration to whether volume can be reduced by undertaking keyword searching. Service providers recommend a list of key words between 30 and 50 terms long - this will usually result in minimal retention of irrelevant material. Don't forget to familiarize yourself with code names; nick names and common abbreviations. Speak to your client and get as much information as possible on the relevant technical terminology. It is also very important that you check your list of keywords with your experts: their input is invaluable. Key word searching can also be used to filter out some privileged documents (use the word "privileged" and/or the names of in-house counsel). It is worth running any key word searches on a sample of data and analysing the material that has been discarded. If it includes relevant material you will need to refine your key words. Doing this provides comfort and trust in the search and will assist you in justifying the stance you have taken if it is challenged. Nevertheless, key words can be a "blunt instrument" for filtering documents so should be used with care. Make sure your documents are OCR-ed[2] before you run any keyword filtering. Keyword filtering/searching is provided by all vendors to varying degrees of usefulness. You should ensure that you check out the functionality during your assessment of the supplier/products.

2.2.3   **Clustering** - uses the same technology as concept searching to group like documents together and present graphically. Allows for further review than traditional linear method. *Example provided by FTI Consulting's Attenex product and by DocuMatrix.*

2.2.4   **Email threading** – allows you to reconstruct email threads to identify conversations of key custodians. *Examples from various tools.*

2.3   N.B. Before you undertake any of the above, you should try to obtain the other side's buy-in.

3   BEST PRACTICE

3.1   Ensure that:

(i)   You keep a detailed record of each process applied to your documentation from collection onwards so as to provide a suitable audit trail for what process has been applied to each document, including a detailed record of the methodology and logic used to remove any documents from the pool;

---

[2] OCR stands for optical character recognition. If, for example, a PDF is simply scanned, an image of the document is produced. However, it is sometimes possible to convert documentation such as a PDF into a file that has OCR, and is searchable to a greater or lesser extent.

(ii)   Any documents which are excluded from the pool of potentially disclosable documents will not be discarded;

(iii)   Keep a separate file of copy correspondence about disclosure; one section for correspondence with the client, another for correspondence with your opponents.