eDisclosure and harvesting the documents in construction cases

First published in the Chartered Institution of Civil Engineering Surveyors (ICES) Construction Law Review, July 2015.

In this article Simon Tolson explains how the use of technology may assist in reducing the cost of litigation/arbitration by expediting and improving the process of disclosure of documents.

Harvesting:

In the bad old days, the full scope and complexity of an eDisclosure exercise often only became apparent later in proceedings, and usually only after the exercise had already begun. As a result, collecting data and assessing the costs of eDisclosure in the past were not often considered until relatively late in the case (although the pre-action protocol provided an early focus).

For the disclosing party, the steps involved in a possible eDisclosure exercise include: considering how to preserve and use documents; a scoping exercise to assess which documents are involved; considering what to disclose; whether special software is required; identifying software/vendors; and discussion with the other party.

The first step in eDisclosure[1] terms is 'harvesting' the documents; namely identifying and collecting the data. This can be a mammoth task. Around 18 years ago, increasing numbers of businesses and individuals went over to creating, exchanging and storing their documentation and communicating with each other entirely by electronic means. The end product is that a colossal volume of information is now created, exchanged and stored only electronically. Email communication, documents, spreadsheets, programs, modelling, whether financial, engineering (BIM) or risk management), accounting, QA, drawing registers, and ever-increasing other forms of ESI now form the bulk of the documentation held by companies, other enterprises and individuals who become involved in litigation and arbitration.

The sheer volume of ESI can be problematic, and Lord Justice Jackson captured this predicament in a speech in November 2011 while giving the seventh lecture in the programme for implementation of his reforms as recommended in his *Civil Litigation Costs Review Final Report*[2]. He said (paragraph 2.2):

> *"The problem: Even in medium sized actions where all the documents are in paper form, disclosure can be a major exercise which generates disproportionate costs. It can also result in a formidable bundle, most of which is never looked at during the trial. In larger actions where the relevant documents are electronic, the problem is multiplied many times over."*

Lord Justice Jackson also stressed that legal professionals now regularly rely on third-party software consultants, who may be experts in their own particular software but may not understand the needs of a particular case, at least not unless embedded with the case team. He commented that:

> "[Consultants] *understand their own software systems, but it is the solicitors and counsel*

[1] eDisclosure is the process of identifying, preserving, collecting, filtering, reviewing and disclosing electronically stored information (ESI). This includes information stored on personal computers, iPads/mobile devices/PDAs, mobile phones and USB memory sticks, as well as email, documents and calendar files.

[2] January 2010.

> *involved who best understand the case … Disclosure is not an activity which can be outsourced in its entirety to external consultants. No existing software programme is capable of achieving standard discovery."*

When Jackson was saying these things big bang had already hit. Electronic disclosure in civil cases was introduced by the practice direction to the Civil Procedure Rules (CPR 31 — Disclosure and Inspection of documents) in October 2010. From then on, in any instance where documents relevant to a case are stored electronically, the parties have to consider and discuss how disclosure should be carried out at an early stage, and all relevant documents must be preserved from the time when court action was first contemplated.

The problem has been that until recently (and the TeCSA eDisclosure protocol[3]) relatively few solicitors and even fewer barristers really understood how to undertake eDisclosure in an effective way.

Go reap

So, at the starting grid of eDisclosure, where do you find 'the papers' — the ESI — to harvest? There are numerous other sources of documents to consider when deciding where to look for documents. It should be borne in mind that some custodians (senior employees, engineers, planners, project managers, architects and decision makers) are more important than others and their communications and inboxes may require much closer scrutiny.

The original file structure should wherever possible be retained when electronic material is being investigated or collated and there must be an understanding of what document management systems were in use at the time, how they were updated, and where they exist now (company acquisitions and mergers complicate matters). It will not suffice merely to look at paper files or email account inboxes.

Today documents may be found on PDAs like Blackberries and iPhones, held in sound files, on backup tapes, and uploaded in the cloud and in web-based email. They may also appear in accounts such as Yahoo, Google and Bing mail, and on mobile phones, social networking sites, Snapchat or Instagram. Remember too that many individuals have more than one email account — business, personal, webmail. LinkedIn is used, as is Twitter. It is also vitally important to investigate data migration, and make proactive enquiry of the IT department, professional staff and HR department. Consider also the data retention policy, and what happened to any old laptops.

It is important that diverse material is considered and if one party can demonstrate that the material is or is likely to be of relevance on a certain platform to the issues in question in the action they should take steps to ensure it is collected.

Indeed paragraph 4.1 of the Protocol requires each party to keep a detailed record of each process applied to its documentation from identification and collection onwards so as to provide a suitable audit trail for what process has been applied to each category of document, including a detailed record of the methodology and logic used to remove any documents from the collection.

Appendix 1 to the TeCSA eDisclosure protocol provides this helpful reference point for locating and identifying the nature of documents and key custodians:

- *[Identify* **locations** *of categories of documents and key custodians of documents]*

[3] http://www.tecsa.org.uk/e-disclosure.

- *[Identify any categories of documents which are* ***located outside the jurisdiction*** *of England and Wales]*
- *[Identify any categories of documents which are* ***not reasonably accessible****]*
- *[Identify any categories of documents* ***which may no longer exist****]*
- *[Identify any categories of documents in native format which were created* ***using relatively unusual software*** *(e.g. Primavera, Micro station, Microsoft Projects, AutoCAD, the BIM software or any bespoke software)]*
- *[Identify any documents that* ***cannot be collected in native format****]*
- *[Identify any* ***documents/locations/custodians*** *which have not been collected but which are* ***subject to further investigation****] [emphasis added]*

Mechanism and recall

With data collection and harvesting, metadata (literally data about data[4]) can easily be unintentionally altered, or adulterated, by the very act of collection, which in some circumstances can have a detrimental effect on the document's evidential integrity and its age profile.

File system metadata created by programs such as Word and Excel can include the following invaluable information: file name, original author, information regarding by whom and when revisions were made, number of pages, number of characters, file size, date created, date modified and date printed. Email metadata can provide additional information, including the real author, sender's domain, the route a message has travelled over the Internet and where delays may have occurred between sending and receipt. All this is useful information.

What is more, ESI can be moved about nationally and internationally, indiscriminately, and at whirlwind speed, but parties to litigation and arbitration often do not have a clue as to how much ESI they have or where it is exactly. It may be on personal computers or backup tapes. Equally, often the parties do not know where to begin their searches. In the case of email, for example, the relevant servers are often not in their possession and may not even be in the jurisdiction; they may for example be in a datacentre abroad.

Limiting the flood

Here I look at custodians, keywords and date ranges.

Initial harvesting of documents by reference to users, date ranges and keywords is a common approach and filters can be used to (i) exclude irrelevant documents or (ii) help identify disclosable documentation within the wider pool of documentation extracted.

The purpose of keywords in the field of eDisclosure is to assist the document review; first, by narrowing down the dataset to be reviewed, and second, to find the specific piece of information sought. There is nothing 'wrong' with keyword searches as such. Used as a tool to locate relevant material, they are readily comprehensible, transparent and efficient to implement. However, they are a rather blunt tool. If the keyword list is focused too narrowly, highly relevant, disclosable documents will fall through the net; if the list is drawn too widely, then searches will pick up acres of irrelevant material.

[4] See Practice Direction (PD) 31B.5(7). Where copies of disclosed documents are provided in native format, some metadata will be disclosed with each document. In the context of UK civil litigation, a party requesting disclosure of additional metadata, or forensic image copies of disclosed documents (e.g. in relation to a dispute over authenticity), must demonstrate that the relevance and materiality of the requested metadata justify the cost and burden of producing it (PD 31B.28).

A keyword search of ESI is literally a search carried out by using specified keywords, to ascertain which documents are produced (or 'harvested', to use a favourite eDisclosure word). Keyword searches can help to reduce the disclosure to manageable amounts for human review where that is appropriate.

The loss of native structures

One of the other things that must be understood is that the eyes may see but the brain may not deduce what it should because of the way eDisclosure presents the data from the harvest to a reviewer.

Documents rarely exist in a vacuum on their own; they almost always belong to some sort of group. Group membership may be intrinsic to the documents themselves (e.g. a set of board minutes; a chain of email correspondence; weekly progress reports); alternatively, it may exist only in the context in which the documents are saved (e.g. a central archive of project documentation). In either case, the context can give the document meaning; conversely, in the absence of the context, it can be difficult or even impossible to understand the significance of an individual document. A receiving party reviewing the other side's disclosure will almost certainly want to review board minutes as a series, email correspondence in chains, and project documents as a set. Unfortunately, with electronic disclosure, this is often impractical and sometimes impossible, so if possible try and agree a protocol where meeting minutes classes are grouped.

The problem can be exemplified by considering how to identify a complete set of board minutes from an eDisclosure database. Searching for all MS Word documents containing the phrase 'Board Minutes' (by date) would probably be a good start. But such a search is highly unlikely to throw up a clean and complete set of minutes. First, there are bound to be false positives such as references to board minutes in other documents or even (depending on the sophistication of the search engine) documents that happen to have the words 'board' and 'minutes' in them. Depending on the de-duplication regime, there may be multiple copies. In addition, the search will inevitably throw up early drafts of some of the minutes, and it might be very difficult to distinguish drafts from final versions. Moreover, one or two sets of minutes will probably be missed because of a typographical error in the title.

However, if the reviewer is looking for a specific 'needle in a haystack' among the dataset, the keywords used should be very specific and narrow, and can include, for example, the custodian's name, specific date and the subject matter of an email. This technique can be employed in the later stages of the investigation, when the document set has already been narrowed down.

As regards the searches themselves, the simplest form of keyword search is what is known as 'Boolean logic searching'. A reviewer will use a Boolean operator such as 'AND', 'OR' and 'NOT' or 'AND NOT' to link keywords (e.g. 'water' AND 'melon' NOT 'pear'). This, in turn, will only produce hits that match the Boolean search terms. In this example, the only results produced will be those that contain the words 'water' and 'melon'. This is particularly useful when trying to find specific documents.

Words can also be searched based on their proximity to one another. For example, a reviewer can request that the only results returned are those that contain the word 'diesel' w/3 (within three words) of 'UPS'. This can be extremely helpful in narrowing down a data set.

Using these tools can help you find that needle in a haystack in a more efficient manner, both in terms of time and cost. It helps turn a once laborious task into a much more manageable one.

Custodians

A disclosing party (and sometimes a receiving party) will often wish to limit for relevancy reasons the custodians (the keepers of pertinent electronic documents) whose ESI is disclosed to the main players who received 95% of the relevant traffic. There are certain people whose inboxes and outboxes are more likely to contain significant emails than others. Drawing up a list of 'Super Custodians' (i.e. the top 10—12 people) on a major project will invariably capture all the important traffic, even where 200 people were on the job. It is good practice to have sketched out a preliminary list of custodians well before the first Case Management Conference.

Date ranges

The purpose of a date range is to capture data only within a set temporal parameter. So unless the file falls outside a date range of say 3 years from x, and x is agreed, it is a safe backstop.

However, date ranges within which searches should be made and disclosure given have come under judicial scrutiny. In *Digicel (St Lucia) Ltd v Cable and Wireless plc*[5] the court directed that a date range should be varied. In *Picard (Representative of Bernard L. Madoff Investment Securities LLC) v FIM Advisers LLP* the court decided what should be the appropriate date range for disclosure, not the parties.

Summary

eDisclosure can be enormously painful if not handled properly. There is a lot to be learnt about how best to manage the practical and technical complexities of the process consistent with best practice. Information technology is such a central part of commercial life that all parties to litigation need to embrace the requirements it imposes upon the effective conduct of disputes.

Simon Tolson, Chairman of TeCSA and Senior Partner at Fenwick Elliott LLP

stolson@fenwickelliott.com
www.fenwickelliott.com

31 May 2015

[5] [2008] EWHC 2522