Conference

TeCSA/TECBAR

E-Disclosure & E-Working in the TCC

Thursday, 30 June 2011

TeCSA

The Technology and Construction Solicitors Association

TECBAR

Technology and Construction Bar Association

Programme

Time	Duration	Topic	Speaker
4.45	25 mins	Registration; Tea & Coffee	
5.10	5 mins	Welcome	Simon Tolson, TeCSA
5.15	30 mins	The search for the Holy Grail in e-disclosure: Avoiding spam-a-lot	Tom Duncan and Kwadwo Sarkodie, Mayer Brown
5.45	5 mins	Questions from the floor	
5.50	15 mins	A litigator's guide to buying E-disclosure services	Mike Taylor, i Lit Ltd
6.05	15 mins	E-disclosure, Needles & Haystacks: What can go wrong – and how to avoid it	Alex Charlton QC
6.20	5 mins	Questions from the floor	
6.25	15 mins	Report on E-Working: a view from the Bench	Edwards-Stuart J
6.40	15 mins	TCC News	Akenhead J
6.55	5 mins	Questions from the floor and closing remarks	Chantal-Aimée Doerries QC, TECBAR
7.00		Drinks and canapés	

Venue

Berwin Leighton Paisner LLP
Adelaide House
London Bridge
London
EC4R 9HA

Date

Thursday, 30 June 2011

Time

Registration 4.45 – 5.10pm. Drinks and finger buffet from 7pm

The search for the Holy Grail in e-disclosure: avoiding spam-a-lot

Tom Duncan and Kwadwo Sarkodie

"*Judges seek the Holy Grail that will give them just 10 key documents in any case and not require them to find the proverbial needle in a haystack.*" (His Honour Judge Simon Brown QC)

The objective is to achieve:

- Total precision: ensuring all the documents you identify are relevant.
- Total recall: ensuring all the relevant documents are identified.

This presents a number of challenges, but there are practical steps that can be taken.

Identification and collection of documents

Step 1: Ensure the potential sources of the documents are identified and the documents are preserved (Paragraph 7 of Practice Direction 31B).

Step 2: Obtain a good understanding of the client's IT systems and what is and is not available. Questions that need to be considered from an early stage include:

- Who are the custodians of the data?
- Do they have laptops?
- From which period is data required?
- Were public folders used on the project and what systems were used when filing into those public folders?
- What is the client's deletions and archiving policy?
- What back-up tapes are available?

Step 3: Prepare a clear brief for the client to follow when collecting the documents from its systems. This will also assist the process of agreeing the steps with the other side. Make sure complete records are kept of what has been done.

Step 4: Collect the documents in their native form.

Processing the documents

Keyword searches

- Keyword searching is a commonly-used means by which to reduce the data to a more manageable size.

- Solicitors are experienced in devising, agreeing and undertaking keyword searches.

- However, keyword searching has its limitations, and so additional and/or alternative tools and approaches should always be considered.

Other tools

- Paragraph 26 of Practice Direction 31B - "*it will often be insufficient to use simple Keyword Searches or other automated methods of searching alone*".

- Examples of other automated methods:

    - Email threading – All emails which appear wholly in another email are suppressed/deleted, reducing a body of emails to a series of email chains.

    - Clustering – Documents are arranged into groups according to subject-matter, by a computer algorithm which identifies semantically-related words within the documents.

    - Predictive coding – A computer system is fed a sample set of relevant documents, and then identifies documents from within the wider body of data which are similar.

- NB: manual reviews remain important.

Ensuring effective processing

- There will always be the temptation to agree and use those methods with which are simple, tried-and-tested and familiar.

- Some of the alternatives to keyword searching may be more difficult to explain.

- Agreement of alternatives is not always readily achievable because of the litigation mindset, and may require further time and effort.

- However some of the less commonly-used approaches and methods offer the potential to significantly improve the effective and efficient processing of data.

Finding the Holy Grail

The steps that can be taken , include:

- Clients setting up effective document management processes, for example the use of public folders.

- Solicitors engaging with the client from an early stage, for example putting in place a coherent and structured plan to manage the disclosure process.

- Educating all participants about the alternatives to keyword searching and the advantages and disadvantages of the various processing techniques.

- The judiciary encouraging the parties give due consideration to all of the tools and techniques which may be appropriate.

Tom Duncan and Kwadwo Sarkodie are senior associates in Mayer Brown's Construction and Engineering Group in London

Spoilt for Choice

A litigators guide to buying e-disclosure services

A previous version of this article appeared in the September 2010 issue of the Society for Computers and Law Magazine.

A half hour trawl of Google will demonstrate that there are, at least, 50 companies operating in the UK that offer e-disclosure services. How to choose which provider is best for you, best for your client and best for your case is a complex and often confusing process.

At the most basic level external service providers provide essentially the same service, in that they take data and process it against agreed criteria and provide it back to clients in a format which is capable of review. I'm absolutely sure that most service providers would argue that they provide better, faster, safer or additional services to their competitors. However what is certain is that the methodology by which the service is provided varies from provider to provider as does the price, the pricing points, and the assumptions used to arrive at a final cost for the purposes of providing a quote.

Add to this mix the less easily quantifiable elements of, reciprocity, reputation, personal relationships and will on the part of the solicitor and client to assess the market it soon becomes apparent that external service provider selection is rarely empirical.

Your Size Matters

Large law firms sometimes conquer these problems by employing litigation support managers who, amongst other things, spend their time assessing the capabilities of e-disclosure service providers and developing relationships and pricing strategies which suit both the law firm and the service provider.

Mid tier and smaller firms very rarely employ Litigation Support Managers and so are immediately at a disadvantage when it comes to disclosure. These firms need to have in place procedures to allow knowledge learnt from one electronic disclosure exercise to be passed to the next individual or team embarking on a similar exercise. Of particular importance is how to procure the services of an appropriate external service provider.

Know What You Want

When selecting a service provider, firms of all sizes need to be clear about what services they need if they want to be in a position to get realistic quotations and time estimates. Consequently a comprehensive and detailed scoping exercise is very important. This process is made easier by the new ESI Questionnaire, but the questionnaire does not do the work for you! It is important that detailed questions are asked both of the clients IT department and of the clients employees who are involved in disclosure in order to fully understand the totality of what your clients have under their control. From that bedrock of knowledge informed and defensible decisions can be made against the backdrop of the value of the claim about what is reasonable and proportionate to disclose in any given case.

Solicitors also need to know what they themselves have in terms of technology for the review of documents, if you have a review tool available internally can you use it (i.e. is there capacity in terms of server space and technical support) to host the documents in house. Or do you want to utilise an externally hosted service (and if so would you prefer that service to be provided by the same external service provider who does the document processing or an alternative external service provider).

The Process of Selecting a Service Provider

There are various major points which firms would be well advised to pay particular attention to these are;

1. The external service provider's capabilities.
2. The size and experience of the external service provider.
3. The external service providers charging methodology.
4. The external service providers working assumptions.

1. External service provider's Capabilities.

The way in which external service providers process the documents they are provided with can be very different. Law firms should ask their external service providers about their capabilities, including;

1. Where does the processing of data occur? (if it is outside the EEA there could be data protection issues)

2. Does the external service provider use an off the shelf document processing engine or is it an application they have developed themselves? (If it is off the shelf are there any known issues with the product, if it is a proprietary application how has it been tested and benchmarked?)

3. Does the service provider's process utilise lists of "Noise" or "Stop" words and if so are these lists modified depending on the contents of keyword lists?

4. What is their daily document processing capacity? (Not really to assess speed as most service providers will deliver documents to a review tool on a rolling basis faster than they can be reviewed, but to assess the level of sophistication of the organisation)

5. Does the service provider have their own data collection and forensic capability or do they subcontract those elements to a third party? (If they do use sub-contractors who are they and what are their qualifications?)

6. What document types, if any, is the external service provider unable to process? (there may be specific technical file types in the litigation which the service provider cannot handle)

7. Can the service provider search and host audio files? (Increasingly important as many companies record all incoming calls).

8. Can the service provider deal with foreign language documentation? (If not is that going to be a problem for your case?)

9. Can the external service provider scan, code and OCR paper documents and then add them to the electronic document collection? (If they do use sub-contractors for this work who are they and what are their qualifications?)

10. Does the external service provider have a hosted review tool option?

    If so;

    a. How good does the firm's internet connection need to be?
    b. Do the lawyers firewall settings need to be altered and will this compromise the integrity of the firm's network?
    c. How is the review tool supported?
    d. Can documents be printed from the review tool?
    e. What security measures are there surrounding the review tool?
    f. How fast does the review tool run?
    g. Does the review tool cope with spreadsheets?
    h. Can the review tool carry subjective coding across duplicate documents?
    i. What project audit functions does it have?
    j. Does the tool incorporate "intelligent" or "predictive" review technologies?

11. Can the service provider create load files for the other parties' document review tool? (The ESI questionnaire requires parties to co-operate on the provision of documents to one another)

12. Can the external service provider paginate and print large quantities of documents if required? (It may well be necessary to print large quantities of documents for the court or less sophisticated parties)

13. How does the external service provider usually archive or delete jobs? (Clients will usually want all of their data removing from service providers systems at the end of a job)

14. Can the service provider give you immediate answers to the questions 7,8 & 9 of the ESI Questionnaire?

The capabilities of the external service provider must, to a large degree, match the requirements of the legal team if they are going to consider using the external service provider for the work required. However don't rule out service providers who are a near match but offer great value for money.

2. The size and experience of the external service provider.

The general perception within the legal community is that it is safer to use larger service providers than it is to use smaller less established providers. This may sometimes be true but there will always be a trade off between size and experience and cost. In order to accurately gauge whether or not it is worth taking the "risk" of using smaller service providers law firms should find out;

1. What experience and qualifications do the people working on the electronic documents actually have?

2. What were the sizes (total number of Gigabytes processed and total number of pages of documentation provided for inspection) of the last 3 electronic disclosure projects the external service provider completed?

3. How many people do they employ working directly in Electronic Disclosure?

4. Are they willing to provide references?

Many law firms end up developing very effective working relationships with smaller service providers who they have past experience with and who they trust to do the work on time and within budget.

3. The external service providers charging methodology.

Most external service providers charge use one of two broad approaches;

1. The majority of the overall price is formed by charging a price per Gigabyte of data that is provided by the client for processing and filtering (data "in the top" pricing).

2. The majority of the overall price is formed by charging a price per Gigabyte of data that is passed for review to the client after filtering and processing has taken place (data "from the bottom" pricing).

Both of the methods above actually give a great deal of control to legal teams about the overall cost of their electronic disclosure exercise.

If "in the top" pricing is used then the scoping phase of the process becomes even more important as parties should only be giving the absolute minimum amount of data to their external service provider for processing.

If "from the bottom" pricing is used then particular attention must be paid to development of the data filters to ensure that as few irrelevant documents make it through to the review stage.

Legal teams often prefer "in the top" pricing (even if it proves slightly more expensive) as it provides certainty of cost to their clients. "From the bottom" pricing estimates are only ever best guess quotes (external service providers are often very good at providing that best guess) until the actual data has been filtered and processed.

On top of these processing charges there are always a great deal of peripheral costs that soon add up, these are far too numerous to list, and by the time the list was created it would necessarily be out of date but very broadly there will potentially be charges for;

1. Data Collection
2. Data Preparation
3. Data Processing
4. Data Manipulation
5. Data Production

6. Data Archiving

4. The external service providers working assumptions.

This is often overlooked by legal teams that are looking to purchase the services of an external service provider. The temptation is to say, as I did at the beginning of this article, "All of these external service providers provide essentially the same service so we'll just compare bottom line pricing and go with the least expensive one".

This is a mistake because in order to provide a quote service providers have to make certain assumptions about the data and about the filters that will be applied to it;

1. **Amount of data collected.**

   The simplest way for external service providers to bring the bottom line cost of their quotation down is to have low estimates for the likely amount of data collected from each individual who is subject to disclosure. This is applicable whether or not external service providers charge using an "in the top" or a "from the bottom" methodology.

   Legal teams, with very little research, can find out the likely amounts of data that each individual subject to disclosure is likely to have in their possession. IT departments can usually give good estimates of mailbox sizes, file share sizes and personal server space size, it is also quite simple to find out the likely amount of personal data (i.e. non system data) held on portable storage devices and laptops.

   Given the relatively small size of the task of finding this information out it is always best to give assumptions on the amount of data to be provided to external service providers to them rather than let them come up with their own assumptions.

   This is not the end of the story though and legal teams must continue to bear the following two points in mind

2. **Explosion rates.**

   E-mail "container files" are the usual manner in which e-mails are stored and their qualities mean that it is sometimes possible for legal teams to collect their own e-mail data and have a preliminary look at it secure in the knowledge that they are not altering the metadata associated with the e-mail.

   Another property of container files is that they compress the data held within them, and so when the e-mails are removed from the container files the sum of all of the e-mails sizes far exceeds the size of the original container file. The actual rate of compression is not uniform and can vary from no compression at all to up to 10 times compression (or more).

   Most external service providers will charge for the size of the uncompressed (exploded) container file. This means that if a external service provider is charging £500 to process 1 Gigabyte of data and their client provides them with 1Gigabyte of .pst data (Microsoft Outlook's container file) the likelihood is that it will not cost £500 to process but anywhere between £500 and £5000 depending on the rate of compression.

   Clearly external service providers who are looking to lower their overall quote will estimate a lower compression rate (of say 2 times) working in the knowledge that the likely compression rate is going to be higher (more usually 3-3.5 times the size of the container file) but in order to get a look in at the tendering stage they need a low quote and they'll deal with the price increase at a later stage.

It would be unfair to say that this practice is usual, or even widespread, amongst external service providers but in order to ensure that quotations are being compared on a like for like basis it is advisable for legal teams to specify what compression rates that their potential external service providers should use when giving a quotation.

There are some external service providers who do not charge on the exploded size of the file but on the compressed file size.

3. **Filtration rates.**

   Whilst explosion rates are important whether or not the external service provider charges using an "in the top" or a "from the bottom" charging methodology, the rates that are assumed for filtration really only effect the external service providers costs if they charge using a "from the bottom" charging methodology (although these assumptions will affect legal team assumptions about review team time and cost whichever methodology the external service provider uses).

   If a "from the bottom" charging structure is used then the rates of filtration are extremely important to the overall cost. Without testing the filters it is almost impossible to tell what the proportion of documents passed for review will be and so external service providers use their experience to provide a best guess, usually the guess is pretty good, but legal teams do need to ensure that all external service providers are using the same assumptions, because, as with explosion rates, some external service providers will assume a higher rate of filtration to bring the estimated cost down.

   External service providers who charge using a "from the bottom" methodology are particularly prone to very large swings in price when both the assumptions on explosion rates and the assumptions on filtration rates interplay with each other and so sometimes provide a high end quote and a low end quote. This is useful from a transparency perspective (i.e. the external service providers are acknowledging that the assumptions may be wrong and so prices may vary) but not very useful when legal teams go with the overall prices to their client who generally want to know an exact price in order for them to budget appropriately.

Conclusions

Being prepared before you go to external service providers for quotations allows lawyers to take control of the procurement process. Using the ESI questionnaire will help in this preparation.

It is also vital that those purchasing solutions have a good knowledge of the broad picture of the litigation as a whole in order that they keep an eye on the next steps in the litigation. If, for example, a service provider has the ability to speed up the document review process using "intelligent" or "predictive" technologies then do the savings at that stage make it worth using that provider even if that service provider is initially more expensive?

Being proactive and engaged in the process will allow litigators to set the tone of disclosure with the opposing party and to demonstrate the open and co-operative approach they have adopted (should the court ever be interested in the conduct of parties!).

The simple comparison of bottom line costs does not give a like for like comparison and legal teams need to be constantly aware of the various ways in which likely costs can be manipulated by altering basic assumptions.

It must however also be remembered that the vast majority of external service providers wish to offer a great service and real value for money, and have different assumptions behind their pricing because they have different experiences in the marketplace.

Many law firms use the same external service provider time and again, and there are valid reasons for this, not least the personal relationship that builds between lawyer and external service provider which can often benefit legal teams through preferred pricing and service arrangements, as well as growing familiarity with working practices and proprietary tools. However this should never stop legal teams from always making the procurement process competitive and using the purchased solution which most appropriately solves their problem. By keeping external service providers on their toes legal teams will usually obtain better pricing and service than if they use an external service provider out of habit.

E-disclosure, Needles and Haystacks

Alex Charlton and Matthew Lavy

In a recent case in the High Court, concerning the development of embedded safety critical software, a claimant purported to comply with its obligations to give standard disclosure by serving on the defendant 30,000 hard-copy documents and an electronic database containing a further 226,000 documents. There were over 1.8 million pages of documentation. Whilst the case was an important one to the parties and complex in subject matter, the issues were relatively well defined, the amount at stake relatively modest (less than £10 million) and there were no aggravating features such as allegations of deceit or fraud.

As readers will be aware, standard disclosure under CPR, r 31.6 requires a party to disclose the documents it relies upon, documents that support another party's case and documents which are adverse to its or another party's case. The concept of standard disclosure was intended to cure the pernicious ill perceived by Lord Woolf to afflict litigation at the turn of the last century, namely, the production of mountains of irrelevant paper. It would be reasonable to suppose that the intention behind Standard Disclosure was to focus the parties' efforts on finding documents central to the issues in dispute and not those documents, in an IT context, that relate to the ebb and flow of project life as a whole.

Every Needle with Free Haystack

In representing the defendant in the case referred to above, it was readily apparent to us and our instructing solicitors that the database that had been provided to the defendant contained thousands of documents that had nothing to do with the issues in dispute and indeed nothing to do with the project. There were thousands of duplicate documents notwithstanding the fact that electronic de-duplication had apparently taken place. There was no useful structure to the database and no folders of e-mails by individual (something that defendant thought had been agreed). We believed that the defendant was facing a needle-in-haystack situation and applied to the court for relief.

Evidence was filed by the claimant which explained how disclosure had been managed. It was an epic tale. The claimant disclosed that document consultants had been retained for the purpose of 'assisting' with the discharge of its disclosure obligations. The document consultants had 'harvested' documents from a large number of servers and tape backups (not just project servers). There can be no doubt that the ambit of the search was wide. The harvesting technique was simple: unless the file fell outside a date range (5 years) or exceeded a certain size (denoting it was code), everything was to be included. The harvest was rich. Over 8 million documents were collated.

The claimant's explanation of how it identified documents falling within standard disclosure from its initial 8 million document harvest was highly controversial. The claimant had simply produced a set of key word filters and the document consultants had applied those filters to the harvest. Unfortunately, the key word filters failed to reduce the number of documents to a level that the claimant deemed to be appropriate for disclosure. Therefore, on advice from the document consultants, the number of key words in the filter was reduced from 333 to 133. The satisfying but perhaps unsurprising result was that the number of documents captured by the filter was also reduced, to a modest 226,000 documents. Some checks were carried out by solicitors for privileged material and the database was then disclosed. It was noteworthy that the costs paid to the consultants for their document handling exceeded the costs incurred by the solicitors by a factor of about five.

At the hearing of its application, the defendant argued that, on a proper construction of CPR Rule, r31.6, it was impermissible to disclose such vast amounts of irrelevant documentation. Specifically, the defendant submitted that by serving many more documents than those falling within the scope of sub-paragraphs (a) and (b) of the rule, the claimant had failed to comply with the obligation to 'disclose *only*' documents falling within those sub-sections as required by the wording of the rule. Conversely, the claimant submitted that on its true construction the rule required disclosure of *at least* those documents falling within the sub-sections but placed no prohibition on disclosure of further documents that fell outside their scope. In the absence of supporting authority and evident judicial hostility, the claimant did not pursue this line of argument and the judge did not have to decide the point.

Behind the argument over the construction of r31.6 lay a fundamental difference of opinion between the parties as to the efficacy of e-disclosure software. As far as the claimant was concerned, the software was an incredibly powerful tool that had transformed the nature of the disclosure process. After all, using this tool, the claimant had managed to harvest millions of documents, run a scythe through them to find those likely to be relevant to the dispute (by the simple application of keyword searches) and had provided the defendant with a convenient, searchable database of material. It did not matter how large the database was or how many irrelevant documents it contained; documents of interest could easily be found by unleashing the power of the search tools. As Leading Counsel for the Claimant put it:

'*In no sense whatsoever is it the case that the defendants have to look at any significant quantity of irrelevant documentation. They have been provided – if they choose to take it up – … with a system of locating documents which a few years ago I suspect many of us would have given our eye teeth to have. From the picture that one is being given by the defendants one is imagining rows and rows of Lever Arch files stuffed largely with irrelevant documents and someone having to wade through them. In the old days some cases were like that, but this could not be further removed from it. This is an impressive and incredibly helpful tool that is being made available to the defendants to allow them to search within seconds or minutes for documents.*'

Unfortunately, the defendant and its lawyers did not see it that way. What we saw was an enormous electronic database with no useful structure, no order and, in respect of a very large number of the documents, no reliable information as to authorship, ownership or purpose. Of course, the defendant could run searches against the database to identify documents falling with a certain date range or containing certain keywords. However, it did not have the claimant's knowledge of the contents of the database; therefore, when constructing its searches, it never quite knew what it was looking for. Sophisticated and unsophisticated searches were tried and *some* relevant material was found. However, the defendant could not be sure that it had found all (or even most) of the relevant material that the database contained. Searches generally returned hundreds or thousands of documents all of which needed to be reviewed manually for relevance. Short of reviewing every document in the database, it was not possible to understand what documents were missing or what documents had been disclosed. The short point was that there was no search that could be devised that would enable the defendant to identify all documents falling within r31.6(a) and (b), and *only* those documents. Had it been possible to formulate such a search or searches, the claimant, who it must be assumed knew what documents were likely to be relevant, would surely have done so.

What had occurred was a negation of the fundamental principles of disclosure. The defendant did not object to electronic disclosure as such; it certainly appreciated that a searchable electronic database was far better than a mound of lever-arch files. But the claimant, it said, had not *supplemented* a normal disclosure process with a sophisticated electronic search tool; rather it had *substituted* one with the other.

As noted above, the claimant did not pursue the construction point and the judge did not need to decide it. The judge's solution to the practical problem faced by the defendant was to order the claimant to apply a narrower *agreed* set of keyword filters to the database to shrink its size to more manageable proportions and to try to minimise the number of documents not falling within r31.6(a) or (b). This order was predicated on the basis that word searching could in principle provide an imperfect but workable remedy.

The claimant complied with the order. The result was a database of approximately 115,000 documents. After a number of abortive attempts to navigate the material using search tools alone, the defendant ultimately decided to undertake a time-consuming and expensive comprehensive review exercise, gathering a team of people to look at each of the 115,000 documents in turn. The case settled prior to trial, and the question of who should bear the costs of the defendant's costly review exercise was never argued in court. From the defendant's perspective, the claimant had failed to comply with its obligation to provide standard disclosure and the court's solution failed to answer the point.

Conflicting Disclosure Interests

This case, in a nutshell, illustrates a common issue that arises during e-disclosure exercises, namely a misalignment of interests and understanding between disclosing and receiving parties. For a disclosing party, e-disclosure is almost paradise compared with its paper-based counterpart. Data gathering (from paper and electronic sources) becomes a readily-outsourced mechanical exercise; creating a list and locating potentially privileged documents becomes a relatively painless task involving compiling lists of keywords and typing them into a search engine; inspection involves handing over a few DVDs or (more commonly) providing the other side with access to an online database. Furthermore, with its intimate knowledge of the documents, the disclosing party is usually able to find documents within its own database quickly and easily. Given their typical experience of e-disclosure exercises, it is hardly surprising that disclosing parties are frequently enthusiastic about the power of e-disclosure software and the benefits that it can bring to litigation.

Unfortunately, while e-disclosure can be relative paradise for the disclosing party, all too often it is a nightmare for the receiving party. Leading Counsel for the Claimant in our case was not wrong to describe the e-disclosure facility as an '*impressive and incredibly helpful tool*'. However, what the claimant's side appeared not to appreciate – and perhaps *could* not appreciate given their positive experience of the tool – was that in the absence of top-down knowledge as to what documentation exists in a database and how it is structured, a search tool is a very blunt instrument. A receiving party wanting properly to understand the scope and content of its opponent's disclosure cannot rely on search tools alone. It must either undertake a systematic and comprehensive review exercise of the whole database, or plough on with the litigation having no idea of the scope of disclosure and knowing full well that highly relevant material is likely to remain undiscovered to the end. E-disclosure databases tend to contain much more material than their paper counterparts; a systematic trawl is invariably a costly exercise.

The problems so often faced by the receiving party of an e-disclosure database stem from two characteristics of typical electronic disclosure exercises: (a) the bottom-up way in which material is typically selected for disclosure and (b) the substitution of filing system structures with meta-data and search tools. Our aim in the remainder of this article is to explain how this happens, why it matters and what can be done about it.

Problems

Selection of material (or 'curse of the keywords')

One of the first (and most fundamental) steps for the disclosing party in any disclosure exercise is to devise a plan of action for undertaking the document search. In a traditional, paper-based exercise, the starting point will tend to be a conference with the client to determine who was involved in the factual background to the dispute, what documentation they created or acquired, and what they did with it. From this information, a plan will be devised to gather material likely to be relevant, to review it and to select those documents requiring disclosure. A vital characteristic of this approach is that it is top-down: it is guided and informed chiefly by the client's knowledge of the factual matrix.

The starting point in an e-disclosure exercise will typically also be a conference with the client. However, the individuals whose knowledge is sought will not be the actors in the dispute; rather, they will be members of the IT department. The subject matter of the conference will not be the factual background but computers: how many servers does the client have? How are they archived? For how long are e-mails stored? Once these questions are answered, harvesting software will be unleashed, sweeping up and cataloguing every document created or modified within a specified date range. When the harvest is over, the crops will be subjected to an avalanche of keyword searches. If a keyword matches, a document will be flagged for disclosure; if no keywords match it will be discarded. A reasonable search for documents likely to be relevant to the dispute is substituted with an exhaustive search followed by a keyword filtering process.

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 blunt tool. If the keyword list is focussed too narrowly, highly relevant, disclosable documents will fall through the net; if the list is drawn too widely, the searches will pick up swathes of irrelevant material. Typically, a (perfectly proper) decision is made to err on the side of caution and to cast the net widely, accepting the inclusion of irrelevant material as an undesirable but unavoidable side-effect. However, even with such a cautious approach, relevant material will be missed. A typical candidate for omission is informal email correspondence which (for example) does not refer to a project by name. Such informal correspondence can be highly relevant and illuminating; certainly it can be the life-blood of the cross-examiner. It should certainly be disclosed.

In short, an e-disclosure database will typically contain much more material than its paper counterpart (due to the volume of material on corporate computer systems and the ease with which harvesting can be achieved), a much higher proportion of its contents will be irrelevant, and yet highly relevant material will probably be missing.

This problem is not an inevitable consequence of the prevalence of electronic data sources in the modern commercial world, nor is it an intrinsic difficulty with electronic disclosure. Rather, the problem occurs because the typical electronic disclosure exercise starts with a bottom-up trawl, not a top-down analysis, and because document review by the disclosing party is replaced with keyword search.

Structure vs Search

Documents rarely exist in isolation; 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 repository* of project documentation; documents saved in the 'disasters' folder on the Managing Director's computer). 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, e-mail correspondence in chains, project documents as a set. Unfortunately, with electronic disclosure, this is often impractical and sometimes impossible.

In their native habitat, the relationship between documents is usually apparent because most organisations and individual users adhere to reasonably sensible filing structures for documents and (to a lesser extent) e-mail correspondence. For example, someone is likely to be charged with looking after board minutes; they are likely to keep them all together in a particular place. However, after an e-disclosure harvest, this structure – which is inherent in the host filing system but not in individual documents themselves – often vanishes. Within the world of the e-disclosure database, there is no ‘board minutes’ folder. The original structures that illuminated relationships between documents is entirely absent. In its place is a giant electronic warehouse in which each document exists as an isolated entity. Within this warehouse, documents cannot usually be browsed in coherent order (because the original structure is lost). The reviewer’s only tool to impose order on the data is the search facility.

The e-disclosure software market is maturing quickly and many providers now offer highly sophisticated search tools capable of unearthing documents based on multiple complex criteria. As forensic tools, used for example to evidence a witness’s involvement in a specific activity at a specific time, these search facilities are a marvellous thing. However, they are no substitute for structure. They unearth isolated documents but they rarely reveal context.

The problem can be illustrated by considering how to identify a complete set of board minutes from an e-disclosure database. Searching for all MS Word documents containing the phrase ‘Board Minutes’ (ordered 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 applied to the database, 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.

A slightly more sophisticated approach to searching might be to note that the custodian of the board minutes adopted a consistent file naming convention and to execute a search designed to capture all files whose name matches the relevant pattern. However, just as with the typographical error in the title, at least one set of minutes is bound to have a filename that fails to adhere to convention. Perhaps the secretary was absent one week and the temp indolent. Moreover, the efficacy of this approach is highly dependent on the expressiveness of the search language – and in this regard, some software is much better than others. The approach could also be thwarted by over-zealous de-duplication.

In reality, the basic board minute problem in this illustration is not a severe one. A search for the phrase ‘board minutes’ will find most of them; the gaps (if any) can be identified and filled by one or more broader keyword searches (e.g. ‘min*’); the number of hits can be controlled by focusing the broader searches over a narrow date range. However, the process will not necessarily be smooth and it will certainly be unduly time-consuming given the typical size of an e-disclosure database. With more complicated real-world examples, the issue becomes more real and acute. For example, once a complete set of board minutes has been located, there will usually be a need to find papers presented at each of the meetings. Such papers are often significantly more enlightening than the minutes themselves. Whereas in a paper-based disclosure exercise you would expect these to be disclosed alongside the relevant minutes, in an e-disclosure context they will almost certainly not be. Finding them can be very difficult indeed (unless they happen to contain the phrase 'board minutes'). Likewise, typical needs such as locating chains of e-mail correspondence (in the absence of a constant 'Subject' line) or finding all documents worked on by a particular person in a particular date range can be complex and time-consuming to fulfil. Using search tools to answer nebulous but often vital questions such as “is there a gap in the disclosure?” takes a lot of skill, a lot of time and a fair amount of luck. Search tools are no substitute for structure and top-down knowledge.

Metadata

The impotence of search tools can be ameliorated by metadata, but only to a limited extent.

It is important to distinguish between two types of metadata. First, there is data deliberately associated with a document by a (usually human) agent ('type 1 metadata'). This data is sometimes stored in the document itself (e.g. the 'Title' field of an MS Word document) but more usually stored in an external database (e.g. the 'Client' and 'Matter' fields in a solicitor's document management system). Type 1 metadata is meaningful, usually reliable and is conceptually equivalent to the structural information so often lost in the data harvesting process. Where available, it *must* be imported into an e-disclosure database. Good quality Type 1 metadata can extinguish most of the problems referred to earlier in this article.

The bad news is that Type 1 metadata does not often exist. Most talk of 'metadata' in an e-disclosure context refers to the second type of metadata. This second type ('Type 2 metadata') is normally not deliberately associated with a document by a human agent; rather it is generated automatically by system or application software. It includes document attributes such as file type, creation date, modification dates and author details. Whilst this metadata is undoubtedly useful, it is notoriously misleading. For example, the 'author' field will often not contain the name of the person who created the document but the name of the person logged into the workstation where the document was last edited (or perhaps where the template was first created). Likewise, the creation date is likely to refer to a template rather than the document itself; the last modified date could relate to the date when someone last opened the document to *read* it, made no changes but absent-mindedly saved it nonetheless. Moreover, metadata dates are only as correct as the clock on the computer on which they are recorded.

This article is not the place for a full analysis of the various pitfalls of relying on metadata. For now, it is sufficient to note that Type 2 metadata can be unreliable, is not the product of deliberate human intervention, and is no substitute for structure and top-down knowledge. Notwithstanding these pitfalls, metadata is an essential part of any document and, potentially, a good source of information. It is surprising therefore to find that the current Commercial Court guide (and the Technology and Construction Court guide) states that generally metadata '*is unlikely to be relevant*' and (by implication) need not be disclosed.

Does any of this matter?

If you are reading this article with a receiving party's mindset, the reason why these problems matter are obvious. E-disclosure databases are typically much more extensive than their paper-based counterparts; thanks to the lack of structure, limitations of search tools and unreliability of metadata, there is often no alternative to reviewing every document in the database or taking a very big risk. The result is massively increased litigation cost.

On the other hand, if you are reading with a disclosing party's mindset, you might consider the woes outlined above would give you something of a strategic advantage: you provide generous disclosure whilst knowing full well that it will occupy your opponents for months and there is a good chance that they will never find any smoking guns.

This latter view is perhaps a dangerous one. Not only is there a serious risk that you may end up paying for the receiving party's expensive review (either because they win at trial or because they successfully persuade a judge that your approach to disclosure was woefully inadequate) but also, if you handle electronic disclosure in the manner described above, you will almost certainly be in breach of your disclosure obligations in at least three ways.

Breach 1

The judge in our case did not have to decide whether CPR, r31.6 limited disclosure to documents that fell within the description of standard disclosure or permitted the disclosure of the entire haystack. The narrow construction, we would suggest, is the correct one for the following reasons.

- The word '*only*' serves no purpose in the wider construction; the rule would have exactly the same meaning without it. As the word is there, it was probably put there for a reason. Not only is the word there, but – as the judge in the High Court case noted – it is placed after the word 'disclose' and not before the word 'requires'.
- One of the principle goals of Lord Woolf's civil procedure reforms was to make the cost of litigation '*more affordable, more predictable, and more proportionate to the value and complexity of individual cases*'. Only the narrower construction is in keeping with this goal. The wider construction protects the disclosing party but not the recipient.
- It could not sensibly be suggested that disclosing 20 relevant documents and an extra few million irrelevant ones for good measure complies with r31.6, but that would be the effect of the former construction.

Breach 2

CPR, r31.10(3) stipulates that a disclosure list '*must identify the documents in a convenient order and manner and as concisely as possible*'. The practice direction expands on this as follows: '*In order to comply with rule 31.10(3) it will normally be necessary to list the documents in date order, to number them consecutively and to give each a concise description*'. Automated lists produced by e-disclosure software (whether hardcopy or virtual) will rarely comply in that:

- the documents will not reliably be in date order (owing to problems with metadata);
- the numbering system is unlikely to be consecutive or sane;
- document descriptions produced from file names and/or metadata may be concise, but will often not be descriptive;
- generally, the list will not be convenient – most likely it will be useless.

Breach 3

Whatever the true construction of r31.6, it is implicit that disclosure is predicated on the parties and/or their legal advisors *reviewing* the documents to determine whether the documents to be disclosed fall within or outside the test. Without some sort of review, a party cannot determine whether a document adversely affects its case or supports that of an opponent. Without undertaking some sort of review, a solicitor or client cannot properly sign a disclosure statement. The Part 31 Practice Direction states only that it '***may*** *be reasonable to search for electronic documents by means of keyword searches*'. It is highly arguable that a review by keyword searches alone is simply not sufficient for the purpose of providing standard disclosure - indeed, it is difficult to understand how a solicitor can advise a client properly unless the documents have been read.

Fixing the problems

The e-disclosure business has expanded rapidly in recent years. It is likely to expand further. E-disclosure is here to stay, and with good reason. Handled responsibly, there is no reason why it should not deliver on its promise of promoting efficient, agile and less costly litigation.

Too often, however, e-disclosure is not handled properly by the disclosing party. Whether too much faith has been put in the technology or whether the power of the data harvester has made disclosing parties lazy, the burden of disclosure has effectively shifted from the disclosing party to the receiving party. The burden is a heavy one, not only in terms of cost but also in terms of litigation prejudice. There is a good reason why the CPR places the disclosure burden on the disclosing party and requires the completion of a disclosure statement: it is because the disclosing party has (or at least should have) the requisite structural knowledge – it knows what it has, where its documents are and what their significance to the litigation might be.

There are two potential solutions to the problems highlighted in this article. First, there is a technical solution. Over time, e-disclosure software will improve. The '*impressive and incredibly helpful tool*' will become more so. It should be possible reliably to maintain structure during a data harvest and navigate that structure in a convenient way. Moreover, it should be possible to replace missing structure by supplementing the blunt tool of keyword and metadata searches with more sophisticated searches that combine rule-based and statistical approaches to document matching.

Secondly, and more immediately, there is the procedural solution. Disclosing parties should adopt a more rigorous approach to e-disclosure. The starting point should not be the IT department but the relevant actors in the underlying dispute. These people should be asked what they produced, what they received and what they did with it. Only if people with the relevant knowledge genuinely cannot be found should a bottom-up harvest be considered, and only as a last resort. Wherever possible, structure should be maintained as a database is built. If the e-disclosure software cannot handle this adequately, separate disclosure databases should be maintained for, say, project files and e-mail correspondence. Sets of minutes should be imported in sequence and (to the extent possible given technical constraints) linked together. Ideally, a database of project files (where they exist) should be agreed between the parties. Ultimately, large organisations should have disclosure obligations in mind when building IT architectures from the ground up and should make provision for the creation and storage of Type 1 metadata. As soon as the prospect of litigation arises, immediate steps should be taken not only to safeguard relevant documents but also to ensure that employees with the requisite knowledge of those documents remain with the company.

CPR Part 31 encourages the parties to discuss prior to the first CMC how e-disclosure is to be approached and limited. These discussions should be real and pragmatic rather than merely lip service to the procedure. For example, in some disputes, it may be more than adequate to restrict disclosure initially to the e-mail folders of a small proportion of those involved in the project, and later to expand the scope of disclosure in a controlled and agreed manner

Recently, there does appear to be a growing appreciation among litigators (who tend to find themselves on 'the other side' of the disclosure wall on a regular basis) that e-disclosure can be enormously painful if not handled properly. The efforts of the LiST group (http://www.listgroup.org) and particularly their publication of the Data Exchange Protocol (Part 1 of which is now in final form; Part 2 was published in draft on 27 April 2007: http://www.listgroup.org/publications.htm) are extremely encouraging. Paragraph 1.2 of the protocol implicitly recognises some of the issues referred to in this article and addresses the core point head on: '*It is a fundamental assumption of this Protocol that parties wishing to receive orderly disclosure in a business-like manner must be prepared to provide orderly disclosure in a business-like manner.*' This is a sentiment that should be thoroughly endorsed.

Alex Charlton is a barrister and member of the IT Group at 4 Pump Court. He is a trained and experienced mediator (WIPO Panel) and a qualified adjudicator.

Matthew Lavy is a barrister and member of the IT Group at 4 Pump Court. Prior to being called to the Bar, he spent time as a technical writer, system administrator and software developer.