Litigation Support SIG – Feedback on CPR Pilot - 23/02/18

The International Legal Technology Association (ILTA) is a premier peer networking organization, providing information to members to maximize the value of technology in support of the legal profession. It was founded in 1980 and runs a worldwide programme of educational events.

The UK chapter of ILTA was formed some 15 years ago, and has a number of Special Interest Groups (SIG's), including one for Litigation Support professionals. This document constitutes the formal feedback from the group on the proposed changes to the eDisclosure Civil Procedure Rules.

The content was derived from a specific meeting held to discuss our response, and has been circulated to members of the Litigation Support SIG to ensure it reflects a common viewpoint. As well as our meeting, various individuals from the SIG attended one or more of the briefing events that have been held over the past few weeks.

Andrew Haslam is currently the lead of the SIG, and has acted as the moderator on these comments. Andrew is currently the UK eDisclosure Project Manager at Squire Patton Boggs, a role he has held for some 18 months. Prior to that he was an independent eDisclosure consultant for over 20 years. Andrew produces an Annual Buyer's Guide to eDisclosure systems, with 6th edition currently under draft. He was a member of the working party that produced the Technology and Construction Court (TCC) eDisclosure protocol, with particular responsibility for drafting the both versions of the Guidance notes for completing the document.

As a group we focused on the specifics of the Disclosure Review Document (DRD) as we felt this was where our day to day experience would be most useful. Overall we were impressed by the document, but we do have a few points for your consideration.

CONFLATING ANALYTICS WITH TAR

First one of definitions. As an industry we use two terms that have specific meanings. We have **Analytics**, that is a suite of technology tools to assist with managing data, including, but not limited to, keywords, clustering, concept searching, near-duplication, email threading and, finally the specific tool of **Technology (or Computer) Assisted Review**; TAR or CAR. TAR is generally taken to mean the process of "training" a computer by a qualified member of the legal team involved in the matter, to the point that the TAR engine can start to identify potentially relevant material with a specified accuracy rate. Indeed it is this definition which is referenced in all three of the eDisclosure precedents we have to date, Pyrrho, BCA trading and (as at last week) Triumph.

You yourselves make this distinction in the final point on page 19 of the DRD, at line (G) which states;

"(G) Any use of clustering, concept searching, e-mail threading, categorisation and any other form of analytics or technology assisted review."

Our problem is with Question 12 of Section 2, reproduced below in italics to provide context

Technology Assisted Review (TAR)

Parties are to consider the use of TAR to assist in the review.

Where parties have considered the use of TAR but decided against it at this stage (particularly where the review universe is in excess of 50,000 documents) they should set out reasoning as to why TAR will not be used.

If the parties are in a position to propose the use of TAR in advance of the CMC, those proposals should be set out in this Section.

From the information given as the various roadshows our group members attended, it seems that the intent of this Section is to determine what tools in the **Analytics** suite will be used, and not just that of **TAR**. It would appear the definition of Analytics has been conflated with the TAR acronym, which has the potential to cause confusion.

We would suggest that TAR is replaced either by the word Analytics, or just by the non-capitalised term technology. A revised Section 12 is shown below for your consideration:

Use of Analytics Tools

Parties are to consider the use of as many as possible of the tools in the Analytics suite available to them, to assist in the review. This might include Technology Assisted Review (TAR).

Parties should identify which Analytics tools they will be using and any configuration applied to those tools.

Where parties have considered the use of TAR but decided against it at this stage (particularly where the review universe is in excess of 50,000 documents) they should set out reasoning as to why TAR will not be used.

If the parties are in a position to propose the use of TAR in advance of the CMC, those proposals should be set out in this Section.

SPECIFIC POINTS ON SECTION 3

Production

Next we looked at Section 3, and in particular some of the points under Paragraph 6 which talks about agreeing aspects of the methodology.

We noted that Point (E) also mentions both Analytics and TAR as distinct subject areas.

On Point (F) we were unclear as to the purpose of the item.

(F) The approach and format for production. This will have an impact on the approach to the review exercise, so parties should endeavour to agree this point at an early stage.

After discussion, we wondered if this was alluding to a possible issue based production in which it was agreed to exchange the issue codes applied to relevant documents? If this is the purpose of this point, it needs to be clarified, if it is not, then what is it trying to achieve?

Approach to de-duplication

We had some feedback on this area, as follows:

"*To the fullest extent practicable, deduplication of the data set should be undertaken prior to giving disclosure of data to the other side*".

This sounds perfectly reasonable, but I'm sure a good deal of deduping is already going on, particularly at the processing stage. What is less clear is the extent to which we will be expected to de-dupe near dupes based on textual analysis. Should we be aiming to review for these purposes (and potentially weed out) anything that is more than 98% similar? More than 95% similar? Or not using this type of de-duping at all?

In our view the technology is not sufficiently reliable for us to feel it appropriate to use it to weed out documents at all.

When it comes to preparing trial bundles, a fairly ruthless approach would be taken in that if there were a number of near dupes then they would almost certainly all be weeded out unless it was clear that any minor differences may actually be of relevance to the matters to be considered at trial. However, this type of approach is not likely to be appropriate at the disclosure stage. You would risk both:

a. Spending inordinate amounts of time manually analysing the differences in documents flagged as near dupes and agonizing over whether they can safely be pulled from the disclosure set; and
b. Getting it wrong, i.e. pulling out documents which you consider to be duplicates on this basis, only for your opponent to object and to say that you should have disclosed and produced them (albeit that in practice an opponent would often not find out).

Our initial view on it all is that we will continue to de-dupe on processing (i.e. de-dupe according to MD5 hash values), but when it comes to near dupes we will aim simply to identify these in our list of documents as being ones which we believe are dupes (perhaps by highlighting them all in a particular colour). We will then continue to produce them.

Whether you feel it is worth adding any comment to the response document I don't know. I can't see a practical way for the pilot documents to give more detailed guidance on the approach to de-duping given that the de-duping technology will continue to evolve, and hopefully improve, and given that what is appropriate (i.e. how far you should go in trying to weed out apparent dupes) will be very fact sensitive depending on the case.

Review and coding email groups

Finally we discussed Points (G), (H) and (I), from which we took an intent to have a hierarchy of requirements

1. Where possible, documents should be produced in Native mode (we agreed with this)
2. Family groups should be kept together (again, we were in broad agreement on this).
3. If documents within a group are withheld, they should be replaced with a placeholder, (here we have substantive issues and disagree).

Our practical objection to placeholders is that we end up with review databases “clogged up” with single page PDF’s bearing a variety of messages. Worse, at times we have “documents” that consist purely of pages of fully redacted text, a pointless waste of time, money and printing ink.

If a document isn’t relevant, don’t produce it. If you want a list of the Privileged documents, ask for it, the data will be in the review system and it’s a matter of minutes to produce a spreadsheet of the key details of those documents.

We don’t need or want placeholders, could the wording in (I) be modified to “*parties might consider the use of placeholders…”*

In discussing this issue, we realised that there were actually two schools of thought on how to approach the coding and production of family groups, that is emails and their attachments. This is one of those seemingly subtle technical points that actually has a major impact upon how the review and tagging work is carried out, and then carries through into the production itself. We have tried to articulate the two approaches below so that you can consider how best to reflect these choices in the DRD. There is no right or wrong approach, a party might adopt both techniques within a matter depending upon the subject material, we mention this because there seemed to be a slight bias within the document to one approach only.

The rest of the document walks through the issues about family groups, and culminates with some proposed wording for the DRD.

In order to explain the differing approaches, we need to delve into a little technical detail. For the vast majority of our practitioners producing emails in Native format means that the whole of the email family is handed over, attachments as well as the email “parent”. If one of the “children” in a family is kept back, then the parent has to be turned into a PDF rendition to avoid inadvertently handing over the child. The significance of this is that every email turned into a PDF, is now a document that can no longer be included in email threading, or identified as a duplicate when exchanged, thus downgrading the use of these tools.

For the purpose of this discussion we are ignoring the children that get produced when emails are processed into any of the litigation support systems, that actually are just the icons/footer messages used in the address block. These are effectively “noise” that we try to strip out before the review teams start work. They are irrelevant, but not coded as such, otherwise every email would fall under the grouping of “Parent Relevant / Child Irrelevant” and be turned into a PDF.

Easy category first of all, all elements in the email are relevant, all elements produced as Native.

We then have two categories where there are different approaches:

- Parent Relevant / Child Irrelevant

- Parent Irrelevant / Child Relevant

Parent Relevant / Child Irrelevant

All practitioners agree that we produce a PDF parent when a child is irrelevant by virtue of Part – Priv or non-relevant and commercial in confidence. The difference in nuance comes when there are non-relevant children, and the difference is down to what should be the prime objective; keeping the parent email as Native, or ensuring you do not inadvertently produce non-relevant commercially sensitive material.

One approach looks as non-relevant and non-sensitive material as “relevant by virtue of attachment” and tells reviewers to code these documents as Relevant so that the whole family group appears as Relevant and the parent email goes out the door as a Native file. The danger here is that, unless this approach is agreed with the opposing party, you can run the risk of being accused of “swamping” the production with irrelevant data. The advantage is that more emails are received as Natives and as such can be processed by analytics tools as emails, not PDF’s.

The opposing view was provided in one of the feedback emails to this document;

- If a parent is relevant but the child is not, then the child should be tagged as irrelevant. We think the overriding objective is better served by weeding out clearly irrelevant documents, rather than by disclosing in Native format.
- We also do not feel it serves the commercial and confidentiality interests of our clients to disclose irrelevant material.
- While under the approach you summarise confidential irrelevant material that is clearly confidential would be withheld, there remains a risk that a client's commercial interests may be compromised by disclosing irrelevant material which is not obviously confidential/ commercially sensitive. Indeed, a review team and even supervising lawyers will not always be well placed to spot irrelevant but potentially commercially sensitive information.

Parent Irrelevant / Child Relevant

Again, two different views. On the one hand is the approach that says you only produce the child document, no need for the irrelevant parent. The other is that the parent email provides context to the child, and therefore you should provide the family by coding the parent as relevant, even when it is not.

As before, the approach needs to be communicated and agreed with the opposing party BEFORE the review work starts, otherwise at worst, you run the risk of having to re-do the review, at best, there is an expensive exchange of correspondence complaining about the disclosure.

We wondered under point (H), there should be wording:

“Parties should describe and agree the coding strategy they will adopt for email families”

ILTA Litigation Support SIG
23 February 2018.