

Clumsy Redaction Can Spell Negligence

By Craig Ball
Law Technology News
February 25, 2008

"The forceps of our minds are clumsy forceps," observed H.G. Wells, "and crush the truth a little in taking hold of it." Clumsier still is a method commonly used to redact information from electronically stored information -- one that so crushes truth, it's alarming anyone defends it, let alone promotes it as a "standard."

I speak of redacting electronic documents by converting them to TIFF images, blacking out privileged and confidential content, then clumsily attempting to recreate electronic searchability by optical character recognition. When applied to spreadsheets and databases, it simply doesn't work. Why, then, are we content to spin invisible cloth rather than acknowledge the emperor's privates are on parade?

Good sense and fair play dictate that redaction methods preserve the integrity of unredacted content and the searchability and usability of the document. Instead, expediency and anxiety drive use of TIFF and OCR for redaction, enabling counsel to cling to familiar, if shopworn, "black line" redaction methods out of fear that privileged contents lurk in some dark digital recess.

To appreciate the problem, consider a complex spreadsheet like those routinely encountered in e-discovery. Spreadsheets are data grids made up of "cells" formed at the intersection of rows and columns. Cells contain hidden formulae entered by the user that generate calculated values seen as numbers in the cell. Formulae are what distinguish a spreadsheet from a word processed table and may be important evidence in that they establish the origins, dependency and sensitivity of the calculated values. Put differently, formulae make the numbers dance. Without them, cell values are runes bereft of rhyme or reason.

With its embedded content, page-defying proportions and dynamic functionality, the exemplar spreadsheet fairly cries out for native production. Alas, it also harbors privileged or confidential content that must be excised.

If the requesting party isn't vigilant, here's how redaction goes wrong:

First, the producing party images the spreadsheet in TIFF format. It sprawls beyond the bounds of an 8 1/2 x 11-inch page, so the data spills confusingly across multiple pages of TIFF images, obscuring column and row relationships. It's a mess.

Second, converting the spreadsheet to TIFF strips away all the underlying formulae, destroying spreadsheet function and undermining a key advantage of native production.

Finally, converting to TIFF means the data is no longer intelligible as data -- i.e., it's not electronically searchable. A TIFF is just a picture -- static ink on a virtual page -- and no more electronically searchable than a Gutenberg Bible.

But it gets worse. To this point, the spreadsheet has been folded across unnatural dimensions, stripped of its usability and rendered electronically unsearchable. Now, the producing party redacts objectionable information like it was any 2D paper document -- by using a drawing utility to black it out or printing it to paper for obliteration by a trusty felt-tip marker!

The spreadsheet's on life support. Seeking to resuscitate its electronic searchability, the producing party administers OCR.

OCR is inherently error-prone, but when the optically recognized data is text, spell-checking corrects egregious recognition errors and restores some of the electronic searchability the federal rules require. When the data is numeric, however, there are no means to spell-check the inevitably myopic OCR. Wrong numbers replace right ones, and the data becomes wholly untrustworthy. By the time the spreadsheet reaches the requesting party, it's a goner:

- Usability: gone;
- Searchability: crippled;
- Integrity: destroyed;
- Content: affirmatively misrepresented.

The operation was a success, but the patient died.

If this is an "industry standard" practice, then we must recall that an entire industry can be negligent. As Judge Learned Hand wrote, "Courts must in the end say what is required; there are precautions so imperative that even their universal disregard will not excuse their omission." *The T.J. Hooper*, 60 F.2d 737 (2dCir. 1932).

Pre-emptively, requesting parties should hone in on how ESI will be redacted, and if flawed redaction techniques will materially impair usability or searchability, they must act swiftly to combat their use and promote alternatives.

Redaction of ESI should be tailored to the nature of the data, using the right tool for the task. Where once native redaction was daunting, now there are reliable, cost-effective techniques for Adobe Systems Inc. PDF and Microsoft Corp. Office documents, including spreadsheets.

For example, [Adobe Acrobat 8.0](#) supports data layer redaction, and the latest release of Microsoft's Office productivity suite stores documents in readily redactable XML formats. TIFF-OCR has its place, but when it's the wrong approach, don't use it. Opt instead for techniques that preserve the intelligibility and integrity of the unredacted content.

Craig Ball, a member of the editorial advisory boards of both Law Technology News and Law.com Legal Technology, is a trial lawyer and computer forensics/EDD special master, based in Austin, Texas.

<http://www.law.com/jsp/legaltechnology/pubArticleLT.jsp?id=1203677136153>