

Page Equivalency and Other Fables

by Craig Ball

[Originally published in Law Technology News, August 2007]

When the parties to a big lawsuit couldn't agree on a vendor to host an electronic document repository, the court appointed me to help. Poring over multimillion dollar bids, I saw the vendors were told to assume that a gigabyte of data equals 22,500 pages. If the dozens of entities involved produced their documents in a mix of .tiff images and native formats—spreadsheets, word processed documents, e-mail, compressed archives, maps, photos, engineering drawings and more—how sensible, I wondered, was it to assume 22,500 pages per gig?

It's comforting to quantify electronically stored information as some number of pieces of paper or bankers' boxes. Paper and lawyers are old friends. But you can't reliably equate a volume of data with a number of pages unless you know the composition of the data. Even then, it's a leap of faith.

I've been railing against page equivalency claims for years because they're so elusive and often abused to misstate the burden and cost of electronic data discovery.

"Your Honor, Megacorp's employees each have 80 GB laptops. That means we will have to review 40 million pages per machine. Converting those pages to .tiff images will cost Megacorp 4 million dollars per laptop."

Nonsense!

If you troll the internet for page equivalency claims, you'll be astounded by how widely they vary, though each is offered with utter certitude. A GB of data is variously equated to an absurd 500 million typewritten pages, a naively accepted 500,000 pages, the popularly cited 75,000 pages and a laggardly 15,000 pages. The other striking aspect of page equivalency claims is that they're blithely accepted by lawyers and judges who wouldn't concede the sky is blue without a supporting string citation.

In testimony before the committee drafting the federal e-discovery rules, ExxonMobil representatives twice asserted that one GB yields 500,000 typewritten pages. The National Conference of Commissioners on Uniform State Laws proposes to include that value in its Uniform Rules Relating to Discovery of Electronically Stored Information. The Conference of Chief Justices cites the same equivalency in its "Guidelines for State Trial Courts Regarding Discovery of Electronically-Stored Information." Scholarly articles and reported decisions pass around the 500,000 pages per GB value like a bad cold.

Yet, 500,000 pages per GB isn't right. It's not even particularly close to right.

Several years ago, my friend Kenneth Withers, now with The Sedona Conference and then e-discovery guru for the Federal Judicial Center, wrote a section of the fourth

edition of the Manual on Complex Litigation that equated a terabyte of data to 500 billion typewritten pages. It was supposed to say million, not billion. Withers, who owned up to the error with his customary grace and candor, has contributed so much wisdom to the bench and bar that he can't be faulted. But the echoes of that innocent thousand-fold miscalculation still reverberate today. Anointed by the prestige of the manual, the 500 billion-page equivalency was embraced as gospel. Even when the value was "corrected" to 500 million pages per terabyte—equal to 500,000 pages per GB—we're still talking an equivalency with all the credibility of an Elvis sighting.

Now, with more e-discovery miles in the rear view mirror, it's clear we've got to look at individual file types and quantities to gauge page equivalency, and there is no reliable rule of thumb geared to how many files of each type a typical user stores. It varies by industry, by user and even by the lifespan of the media and the evolution of particular applications. A reliable page equivalency must be expressed with reference to both the quantity and form of the data, e.g., "a gigabyte of single page .tiff images of 8½"x11" documents scanned at 300 dpi equals approximately 18,000 pages."

Consider the column you're reading. In plain text, it's a file just 5 kilobytes in size and prints as one to two typewritten pages. As a rich text format document, the file quadruples to 20 KB. The same text as a Microsoft Word document is 25 KB. Converted to a .tiff image, it's 123 KB without an accompanying load file. Applying a page equivalency of 500,000 pages per GB, a vendor using per page pricing may quote this column as being anything from one page to as many as 61 pages. Billed by the GB, you'll pay almost five times more for the article as two .tiff pages than as a native Word document. A flawed page equivalency hits the bottom line...hard.

So how many pages are in a gigabyte of data? Lawyers know this answer: *it depends*. To know, perform a data biopsy of representative custodians' collections and *gauge*—don't guess—page volume.