



White Paper

Has Indexing Technology Made *Zubulake* Less Relevant?

ABA Article by: Michael D. Berman, Scott Fischer, and Richard E. Davis

Prior to the December 2006 amendments to the Federal Rules of Civil Procedure, the comprehensive *Zubulake* decisions addressed the production of electronically stored information (ESI) that was not reasonably accessible, specifically, backup tapes, using cost-shifting concepts. The December 2006 amendment to Fed.R.Civ.P. 26(b)(5)(B) subsequently introduced the concept of “not reasonably accessible because of undue burden or cost.” As recently noted in P. Rice, *Electronic Evidence: Law and Practice* (ABA 2d ed. 2008), 34, “[a]s hardware and software change, the currently inaccessible (or the ‘difficult to access’), may become accessible. As a result, the cost-shifting precedent of the past may not be a reliable guide to how courts will rule in the future.”

Currently, at least five procedural devices address the proportionality requirement of the rules: 1) Fed.R.Civ.P. 26(b)(2)(C) (cost-benefit analysis); 2) Fed.R.Civ.P. 26(b)(5)(B) (not reasonably accessible because of undue burden or cost); 3) Fed.R.Civ.P. 26(g)(1)(B)(iii) (certification that discovery is “neither unreasonable nor unduly burdensome or expensive”); 4) Fed.R.Civ.P. 26(c) (protective order against “oppression, or undue burden or expense”); and, 5) Fed.R.Civ.P. 1 (just, speedy, and inexpensive resolution of every action). Stated simply, proportionality means that a litigant need not spend \$750,000 preserving and producing information in a case where the damages claimed are \$75,000. Application of the principle is, however, often complex and fact sensitive.

New high-capacity computer and network data-searching technologies have changed the cost-benefit landscape. The *Zubulake* decisions were issued between 2003 and 2005. While the December 2006 amendments did nothing to undermine the legal analysis of *Zubulake*, technological improvements have occurred that suggest re-thinking the information technology aspect of those decisions, especially as they relate to disaster-recovery backup tapes. In short, if the computerization of industry has created an information overflow problem that was addressed in *Zubulake*, innovations in information technology have begun to solve it.

Ms. Zubulake was employed by UBS Warburg (“UBS”). She contended that she was discriminated against in employment decisions and filed suit. During discovery, UBS produced ESI. Ms. Zubulake, however, had kept copies of certain responsive emails and demonstrated that UBS’s production was incomplete. This led to an analysis of UBS’s backup tapes. By directing restoration of a sample of backup tapes, the court quantified the value of the un-reviewed data residing on the remainder of the tapes, and determined who should bear the substantial costs of restoring, reviewing, and producing data contained on those tapes.

Much of the *Zubulake* court’s factual analysis addresses the problems inherent in reviewing backup tapes. Tapes store data in a linear fashion, which effectively means that entire tapes need to be scanned before responsive data at or near the end of a backup tape can be identified and then restored. In order to verify and extract data from backup tapes, the information must first be restored to the form that it was in prior to being stored on the tape. Once the data is restored to its original native state, the data can be searched for potentially responsive information.

Indexing software has changed the technological underpinnings of *Zubulake*, while leaving its legal analysis untouched. “Indexing” is the process of having software create an “index” of all of the words in a database so that a searcher may rapidly find any information in that database. For example, “Windows Search” is installed on many computers running Windows software and it automatically indexes the computer, permitting instantaneous searches. Indexing software performs the same function on a series of backup tapes, permitting searches and retrieval without the full restoration process described in *Zubulake*. And, there are products that perform similar tasks on active data, by “crawling” through all of the data to create an index that permits rapid searches. As a result, contemporary litigants should exercise caution before assuming that *Zubulake*’s information technology analysis remains applicable in the current environment.

In short, *Zubulake* properly addressed the factual, technological, and legal issues as they existed at that time. In the years following, however, technology has advanced to suggest solutions to at least some of the prob-

lems presented to the litigants in that case. Six years is a very long time in terms of 21st century computer technology and tasks that placed significant operational pressures on corporate organizations in terms of cost and time in 2003 may be relatively simple and inexpensive in 2009.

Backup Tape Technology at the time of *Zubulake* and Today

In *Zubulake III*, the court examined the costs claimed by UBS for restoring and searching five, 60-gigabyte DLT backup tapes. The projected costs, without including infrastructure requirements, were:

Task	Item	Item Cost	Total
Restoration services	Consultant Time	31.5 hours @ \$245	\$7,717.50
Script creation	Consultant Time	6 hours @ \$245	\$1,470.00
Computer use	Machine Time	101.5 hours @ \$18.50	\$1,877.75
Administration fee	Percentage	5% of total cost	\$459.38
TOTAL - restoration for five tapes			\$11,524.63

UBS then extrapolated the cost to the entire 77-tape production, which would result in tape restoration and searching costs of approximately \$166,000. The court held that the data were not reasonably accessible because of undue burden or cost, and ordered partial cost-shifting.

Had the same tapes been presented in 2009, the result may have differed. Newer technology can index, search, extract, and forensically copy selected files from backup tapes. One limitation of that approach is that it cannot index every conceivable tape format and backup system available on the market. However, most of today's standard backup systems are within its purview.

The indexing process avoids many of the costs associated with prior tape-review techniques. For example, if an indexing platform is applied by a consultant, that "vendor" generally performs many or all of the services previously imposed as separate line items. An attorney could retain the consultant, who would either use an existing tape silo or, if none is available, for an additional fee, place the tapes in the vendor's tape library. At that point, they would be indexed and searched by the software and "hits" would be preserved. Actual searching can be performed by the client or counsel and, because the software is capable of forensic preservation of the desired "hits," additional consulting fees may be unnecessary. The following table provides the cost structure in an Index Engines project. In this scenario, the data is stored on the vendor's site and the client may run unlimited searches for 30 days. Unlike the tape restoration option, in which the vendor is given the search terms to run, the indexing option allows attorneys to log on to the system as often as required and run as many searches as are deemed necessary. There is a one time charge of \$1,500 for setup and a monthly charge of \$1,500 for 5-users. This charge is the same regardless of the number of tapes involved. The costs are as follows:

Task	Item	Item Cost	Total
Index, search and extract files from tapes	Tape charge	5 tapes @ \$400	\$2,000.00
Setup data for Searching	One-time charge	\$1,500.00	\$1,500.00
Hosting - Unlimited searching	Monthly Hosting	\$1,500.00 for 5 users	\$1,500.00
TOTAL - file extraction			\$5,000.00

Using modern indexing technology, with the stated assumptions, for 77 tapes, the cost of making the tapes instantly searchable would only be \$38,500.00, instead of \$166,000. There would be additional costs, if, for example, the client lacked a tape silo and wanted the tapes to be continuously searchable, because it would then be necessary to lease a long-term storage device; however, if continued accessibility is not required, absent a client-owned silo, there would be a one-time charge for loading the tapes into a library, the cost of which would be dependent on the number of tapes and duration of the lease.

The cost disparity between restoration and indexing, however, remains substantial and it is even more pronounced because another cost that was required in 2003 - - the infrastructure to hold restored backup data - - has been rendered insignificant due to indexing. In 2003, tape restoration required a major investment in additional storage space in order to hold the restored data. For example, in order to restore the seventy-seven, 60-gigabyte tapes involved in *Zubulake*, and keep them readily accessible, it would be necessary to store 4.6 terabytes of data. Such an environment, would have cost in excess of \$30,000 for the hardware alone. In contrast, with tape-indexing technology, only the index needs to be stored, and that is roughly 8% of the data size. Applying this to the *Zubulake* tapes, only 368 gigabytes of storage, rather than 4.6 terabytes, would be needed. That difference would again significantly reduce the cost of tape review because, instead of requiring the implementation of a server environment to restore the original tapes, the index could be stored on a work station or a \$250 external USB hard drive.

In *Zubulake*, the plaintiff sought substantial damages. The \$38,500 cost of an indexed tape review, even considering the added cost of an indexing platform, would most likely not result in a conclusion that the ESI was not reasonably accessible "because of undue burden or cost."

In fact, the availability of indexing also raises the question of whether data on tape is "not reasonably accessible." Judge Scheindlin provided the classic delineation between accessible and inaccessible data in *Zubulake I*, where the court wrote: "The difference between the two classes is easy to appreciate. Information deemed 'accessible' is stored in a readily usable format. Although the time it takes to actually access the data ranges from milliseconds to days, the data does not need to be restored or otherwise manipulated to be usable. 'Inaccessible' data, on the other hand, is not readily usable. Backup tapes must be restored using a process similar to that previously described... before the data is usable. That makes such data inaccessible." Indexing technology, however, makes data on backup tapes usable without full restoration. Tapes can be indexed and searched in a tape library, without additional demands on infrastructure resources. Therefore, in some circumstances, they may no longer be viewed as inaccessible.

Identification, Preservation, Collection and Preservation

Just as the concept of "undue burden or cost" limits production of data that is not reasonably accessible, "proportionality" limits a litigant's duty vis-à-vis active ESI. See P. Grimm, et al., "Proportionality In The Post-Hoc Analysis Of Pre-Litigation Preservation Decisions," 37 U. Balt. L. Rev. 413 (2008). Imposing and monitoring a litigation hold, including locating, gathering, preserving, and culling information for responsive data, as well as production of the data to the opposing side, can be very costly. Thorough efforts in a complex case could involve: identifying the data in its various locations throughout the network infrastructure; forensically collecting the data; safely copying it to a clean, "read only" location; culling the data by properly-designed and tested keyword searches and date range; determining which fields of metadata to export into a review platform; and, providing the resulting set of information to the attorneys for privilege and responsiveness review in an appropriate format. Because employees should not review any of the potentially relevant data, let alone modify, or delete it, before preservation efforts are implemented, there may be a loss of business during the preparatory period before the litigation hold is implemented, and that may require costly, expedited preservation efforts. The chart below illustrates the "hard costs" for a hypothetical, 100-gigabyte set of data:

Costs of Collecting, Preserving, Culling and Processing for Review

Task	Item	Item Cost	Total Cost
Microsoft Exchange Server - data retrieval	Technician time to perform action	7 hours @ \$300/hour	\$2,100.00
Windows File Server - data retrieval	Technician time to perform action	7 hours @ \$300/hour	\$2,100.00
30 Workstations - data retrieval	Technician time to perform action	30 hours @ \$300 for computers	\$9,000.00
Keyword Searching of 100GB of data	Vendor searching by keyword and date range of the acquired data for responsive text and/or metadata	100GB @ \$50/GB	\$5,000.00
Processing of resulting data for responsiveness and privilege review	Vendor processing and on-line hosting for attorneys	20GB @ \$450/GB	\$18,000.00
TOTAL , file collection and processing before review			\$36,200.00

This example presents a relatively small case. A typical, mid-size corporation may deal with many similarly-sized cases every year. For companies in litigious arenas, there can be many more potential litigation matters each year, and each one may require a process similar to the one described above. Obviously, the cumulative costs can become quite significant.

In the hypothetical described above, the potential litigant will expend \$36,000 simply to prepare the information that needs to be reviewed for a single, anticipated lawsuit. These costs do not include attorney review or production costs and are not transferable to another, separate anticipated lawsuit. In short, they are “project based” charges. While the volume of data is well-recognized as a key cost factor, other factors that need to be considered include the increasing complexity of the data types that comprise the growing data volume, and those factors may increase the hypothetical’s cost.

One alternative is an “enterprise solution” that employs search and indexing technology to change the landscape of preservation and collection. Specifically, the program would “crawl” amongst the data residing on a network, including all the mail servers, file servers, internet and intranet locations, workstations and ancillary storage devices, such as external hard drives and thumb drives. The program would function the way Google, Yahoo, Bing, and other similar search engines crawl and search the internet. And the program would index not only the text of every document, but its metadata as well.

This concept is now a reality, and is used by an increasing number of firms to prepare for litigation, long before it becomes “reasonably anticipated.” This proactive approach means that once the “crawl” is completed, and an index is built, counsel can access a search screen and simply run queries across every indexed file on the network. Because of the high speed nature of the indexes, a query for a series of key words and date ranges will return search results in a matter of seconds. Once the results are displayed, they may be tagged and frozen in place – thereby instantly implementing a litigation hold. The tagging and preserving process takes a few seconds to complete. The documents are then copied to a safe location, where they may be reviewed for privilege and responsiveness.

The key concept of an “enterprise solution” is that it is “proactive” and deploys a re-usable search engine. Once the crawl is complete, the firm’s data is indexed and repeatedly searchable without a new project-based expenditure. Incremental and differential crawls are conducted periodically to ensure that the index is kept up to date and new data is included in each search. Once ESI is indexed, a search can be rapidly run whenever litigation becomes reasonably anticipated. Additionally, organizations may choose to use the technology for safeguarding confidential information and for preventive searches to avoid future problems. All of this may be accomplished without the need for project-based consultants, thereby reducing the cost of collection and preservation.

Unfortunately, the initial cost of deployment of enterprise solutions currently remains high. A typical six-terabyte implementation costs approximately \$250,000 to install; however, for firms faced with multiple annual preservations, or litigation involving substantial sums, the initial investment may be justified. For example, if a corporation had to impose fifty litigation holds similar to the size of the hypothetical case described above, the cost would be \$1,800,000 (50 cases at \$36,000 per case). If, however, an enterprise approach was implemented, the cost would drop to approximately \$5,000 per case (the \$250,000 spent on the software divided by 50 cases), plus some costs for in-house personnel and storage.

As predicted by Mr. Rice’s treatise, because of technological improvements, “the currently inaccessible (or the ‘difficult to access’), may become accessible,” and crawling and indexing technologies may change the technological analysis related to “proportionality” and when ESI is “not reasonably accessible because of undue burden or cost.”