CT Summation

# The New Frontier of Data Sampling

*Are You Prepared to Defend Your Data Retention Policy in a Court of Law?*

White Paper

Published: February 2010

**ABSTRACT**

Over 90% of business communications are now electronic, and the challenges of accessing, storing and maintaining that mass of data is no mean feat. Thanks to new developments in the once-obscure area of law known as electronic discovery, every byte has the potential to hurt an enterprise in surprising—and costly—ways.

Electronic discovery, or e-discovery, will have a direct impact on any business using electronic communications and data storage, and may be involved in litigation. In other words: all business. The rapidly evolving legal requirements of e-discovery represent a perfect storm for data storage and data management in enterprises of all sizes—a storm which IT professionals ignore at their peril.

# Contents


THE NEW FRONTIER OF DATA SAMPLING 1

PRODUCTION IS ONLY THE BEGINNING 1

SEARCH IN E-DISCOVERY 2

SEARCH SAMPLING 2

AN EMERGING INDUSTRY 3



**COPYRIGHT INFORMATION**

© 2010 CT Summation, All rights reserved.

The information contained in this document represents the current view of CT Summation on the issues discussed as of the date of publication. Because CT Summation must respond to changing market conditions, it should not be interpreted to be a commitment on the part of CT Summation, and CT Summation cannot guarantee the accuracy of any information presented after the date of publication. This document is for informational purposes only. CT SUMMATION MAKES NO WARRANTIES, EXPRESS OR IMPLIED, IN THIS DOCUMENT.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording or otherwise) or for any purpose, without the express written permission of CT Summation. CT Summation may have patents, patent applications, trademarks, copyrights or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from CT Summation, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

CT Summation, iBlaze, Blaze, LG, and LG Gold are trademarks of CT Summation in the United States and/or other countries. Microsoft and Windows are registered trademarks of Microsoft Corporation in the United States and/or other countries.

The names of actual companies and products mentioned herein may be the trademarks of their respective owners.

# The New Frontier of Data Sampling

Over 90% of business communications are now electronic, and the challenges of accessing, storing and maintaining that mass of data is no mean feat. Thanks to new developments in the once-obscure area of law known as electronic discovery, every byte has the potential to hurt an enterprise in surprising—and costly—ways. The field also heralds a rapidly growing subspecialty among IT professionals who recognize that proper data maintenance will only become more critical as the law leaps forward to keep pace with technology.

Electronic discovery, or e-discovery, will have a direct impact on any business using electronic communications and data storage, and may be involved in litigation. In other words: all business. The rapidly evolving legal requirements of e-discovery represent a perfect storm for data storage and data management in enterprises of all sizes—a storm which IT professionals ignore at their peril.

# Production is Only the Beginning

Once litigation is initiated, the mountain of data held in company servers must be reliably stored and maintained. Failure to do so may result in criminal prosecution and/or monetary damages not only for the company, but also those responsible for the systems themselves. The court in *In re Napster, Inc. Copyright Litig.* held that while data may go missing or overlooked in large-scale production for trial, "the abject failure to preserve an entire source of relevant evidence is sanctionable conduct." And as financial giant Morgan Stanley found when it failed to comply with its electronic discovery obligations in a 2007 case, such lassitude can be a $1.5 billion dollar mistake. Arthur Riel, the former Executive Director of Morgan Stanley's Law IT Department, was subsequently terminated after he turned up evidence of additional illegal activity by company officers during his production of emails and brought it to the attention of his superiors. The reason for his termination? Reading other employees' email. Riel subsequently sued the company for breach of contract, negligence, fraud and defamation.

While such cases demonstrate the challenges facing firms and government agencies involved in large-scale e-discovery projects, they also shed light on the hurdles faced by litigation software firms and data management professionals. The ability to produce the required data in court is just the beginning.

## Search in e-Discovery

Even for small businesses, producing e-discovery data for a single case can easily involve millions of documents. Everything is fair game, from emails and electronic calendar items to video and sound files, digital voicemail messages, web-based materials, copy machine scans, and CRM data from multiple current and legacy systems. New standards mean that data must not only be effectively searched, but searched in such a way that it is *defensible in a court of law.*

If the opposing parties can agree to a method for searching e-discovery data that the presiding judge finds suitable, the process can be relatively straightforward. However, if that search methodology becomes a point of contention, it can become exponentially more difficult and can make or break a case.

As a result, an entire industry devoted to e-discovery management and search has emerged, as companies often require outside experts to manage the e-discovery process. Services include everything from litigation support software to e-discovery experts who consult with counsel and testify to the validity of search methodologies at trial. But new methods of examining e-discovery data and recent legal decisions may make the process less onerous.

## Search Sampling

Data sampling is most commonly associated with political polling, where the ability to draw big conclusions from relatively small data sets has made the examination of large masses of data far less cumbersome and costly. While there are certainly situations in which human document review teams are the right solution, search and sampling are increasingly becoming indispensable tools, as teams of experts are required to cull hundreds of gigabytes of relevant data.

Sampling refers to the examination of a small, representative sample from a larger pool of information in order to extrapolate the entire pool's content. It is a complex undertaking in any circumstance, but as a tool to examine e-discovery data, sampling is fraught with potential pitfalls. What constitutes a "representative sample" of four million emails? Who determines how, and by whom, that sample will be taken? How will it be extracted, processed and examined? How can an outside observer verify that the selected sample is truly representative of the larger whole?

Sampling in e-discovery has been a way to examine large and complex e-discovery tasks more quickly, cheaply—and often more accurately—than page-by-page review. By employing strategies such as the "iterative feedback loop," in which teams of lawyers and e-discovery experts from opposing sides continually hone the search process in stages, lawyers hope to ensure that client interests are represented.

More recently, in *Victor Stanley Inc. v. Creative Pipe Inc.*, No. MJG-06-2662 (D.Md. May 29, 2008), sampling became a judicially mandated requirement. Judge Paul Grimm held that "[c]ommon sense suggests that even a properly designed and executed keyword search may prove to be over-inclusive or under-inclusive…. The only prudent way to test the reliability of the keyword search is to perform some appropriate sampling of the documents determined to be privileged and those determined not to be in order to arrive at a comfort level that the categories are neither over-inclusive nor under-inclusive."

While many observers worry that this decision might lead to costly teams of experts battling one another in court over the minutiae of varying search methodologies, Judge Grimm's decision specifically suggested ways in which e-discovery costs could be contained: conducting cost-benefit analysis to determine the value of sampling, and encouraging opposing parties to agree upon e-discovery standards (such as the iterative feedback loop model). Judge Grimm predicted that continuing scientific evaluation of search tools would yield court-recognized standards obviating the need for the repeated defense of varying search methodologies.

## An Emerging Industry

Currently no bedrock standards exist for search and sampling, and virtually all related issues are left to the discretion of the presiding judge in a given case. The costs associated with large-scale e-discovery efforts can be staggering, but legions of experts are attempting to develop search standards to significantly reduce the costs and labor associated with e-discovery.

Perhaps most importantly, IT professionals responsible for data storage and management must understand their legal obligations relating to storage and maintainance of data. Today, virtually all businesses are subject to e-discovery requests. As e-discovery standards emerge and the litigation support industry continues to grow, informed IT professionals stand to benefit significantly as they will inevitably play an increasingly critical role in the legal landscape.

---

- TREC Legal Track (http://trec.nist.gov/), a project co-sponsored by the U.S. Department of Commerce's National Institute of Standards and Technology and subagencies of the U.S. Office of the Director of National Intelligence, has been scrutinizing various search techniques using publicly available e-discovery data from the Enron case and *U.S. vs Phillip Morris*.
- The Sedona Conference (http://www.thesedonaconference.org), an annual gathering of e-discovery experts, seeks to examine the larger issues, challenges and liabilities of complex e-discovery tasks.
- Electronic Discovery Resource Model (EDRM) (http://edrm.net), a body of "e-discovery consumers and providers dedicated to reducing the cost, time and manual work associated with e-discovery."
- Discovery Resources (http://www.discoveryresources.org/) provides up-to-date e-discovery news and resources.

**CT Summation**
425 Market Street, 7th Floor
San Francisco, CA 94105
800 735 7866 toll free
415 442 0404 tel
www.ctsummation.com

© 2010, CT Summation, a Wolters Kluwer business. All rights reserved.