SYBASE®
An SAP Company

EDMT Solution Helps Big Litigation Between OnlineNIC and Verizon, Microsoft and Yahoo

USE CASE STUDY

WHAT IS STRUCTURED DATA, RELEVANCE AND MORE

Over the past 20 years, structured data created by business, sales, manufacturing and other operations grew exponentially. Almost all structured data such as sales, accounting, payroll, employee, travel, health and other data is managed by relational databases. eDiscovery, traditionally focuses on unstructured data, email and electronic documents, thus neglecting the fact that structured data in many cases contains critical evidence and potential "smoking guns."

The challenge is that databases are different in every organization. Even the common systems are customized for each customer. Even more problematic is the fact that many databases are purpose-built and fully understood by only a few people within an organization. Often proprietary databases are found in large multinationals where the odds are that only a handful of people on the planet are aware of what they contain.

Though database files are discoverable in electronic form, courts have been reluctant to grant plaintiffs broad access to them for litigation. Courts struggle with information that is not contained in discrete documents. There is little case law regarding authenticating structured data, but there are procedures that can be used to try to verify that information pulled from such sources is complete.

The reason why structured data is neglected during eDiscovery is simple: collecting and analyzing data from structured repositories is a complex process that will overwhelm any lawyer, even those with relational database skills and fluency in the analysis of structured database contents. To make things worse, commonly used eDiscovery tools almost exclusively focus on unstructured data such as emails and documents and, as such, barely support the storage of, much less any analysis of structured data.

Discovery of databases is also expensive, but for different reasons than the discovery of e-mail and other records. In e-mail, much of the cost is in human review to protect privilege when producing a collection of records. With databases, cost overruns are more likely to arise when you are trying to get information out of a system. In many cases, pulling a single record is impossible without preserving the larger data set. Some litigators will ask for the whole database — but what of costs and proportionality if the database has millions or billions of records?

eDiscovery vendors have been slow to respond to this issue. Services have been developed to address the low hanging fruit like e-mail and other file types that are relatively easy to build for, but a reliable process for structured data, one with cross-correlation capabilities does not exist — with one exception — EDMT Solution .

BMMsoft EDMT levers the power of SAP Sybase IQ to enable correlation and joint exploration of structured and unstructured data, together.

WHAT ROLE DOES RELATIONAL, STRUCTURED DATA PLAY IN EDISCOVERY?

The most important manifestation of structured data is in enterprise-level relational database management systems (RDBMS). The relational model, defined by IBM in late 70s and widely accepted in the industry, formally defines schemas of rows and columns that follow strict rules, simple relationships, enabling reliability, predictability and search speed. Two key aspects of RDBMS are ACID rules (Atomicity, Consistency, Isolation and Durability) and SQL (Structured Query Language) which guarantee data quality, reliability, uniform, and predictable results for retrieval and analytics.

Each data point in a relational database has little meaning on its own but offers enormous value when the whole system (tables, rows, columns, indices, rules) is taken together. As an example, consider a fictional company that is sued for race discrimination. Searching data for just a few employees, for example, by race, hiring date, job history, and promotions probably won't tell much about the alleged discrimination. However, searching and analyzing race data of new hires, firing, and promotions have a much likelier chance to reveal behavior and patterns that can become evidence in that case.

HOW TO USE STRUCTURED DATA IN EDISCOVERY ?

Understanding structured data is difficult, especially in a large organization with rich data architecture. But by following certain guidelines, using proper technologies, talking to right people, and engaging lawyers with expertise in structured and unstructured data, organizations have a far better success rate. Specifically:

1. IT has to get involved in the process. IT teams working on OLTP, ERP, payroll and data warehouse systems will be able to point you in the right direction. If your company or client has a data warehouse start there and go over the data map. As databases are very frequently updated, IT should provide you with the archive of old, as well as ongoing data, perhaps even provide you with the near-line version of the database. The ability to archive and preserve or place a “hold” on database records is especially important if you anticipate upcoming litigation.

2. Locate relational data relevant for your case. Review data retention rules and database internals. If your organization lacks thorough documentation, learn what you can from existing sources and work on improving documentation processes going forward together with the records management team.

3. Inquire. The best resource for locating structured data is the people who generate or run applications that generate data, and those who maintain that data. Interview the workers who use your organization's data to uncover sources you might overlook.

4. Make sure that you and the opposing counsel agree on the date ranges, data types, and sources required for your specific suit. Then, build your queries based on the constraints of the case so you extract only the data you need. If the amount of data requiring analysis isn't particularly dynamic, you can export it into a spreadsheet program like Excel. If your structured data collection is large, however, you may need to use a database program with more functionality like Microsoft Access.

5. The most difficult part: cross-correlating structured data with “regular” unstructured data (emails, documents, etc.) used in the litigation. This will be the most difficult part, particularly if data volumes are big and the number of records, particularly database records, grow into millions or billions of items.

WHAT ARE THE PRIMARY ISSUES WHEN COLLECTING AND PRODUCING STRUCTURED DATA?

Collecting and compiling an exhaustive structured data collection has many challenges.

One challenge is collecting data from outdated systems, often called "legacy" systems. The challenge of legacy systems is the lack of intellectual capital available to support them. Finding experts who are well versed enough to get the data out of the legacy system but who can also clearly describe their extraction methods to the court can be both expensive and time consuming, but is necessary when constructing a defensible collection.

A common problem, one that is very frustrating is handling missing or corrupted data. First fix should be locating the original source to make sure the data was correctly exported from the database. If the data source is corrupted, all you can do is to piece it together as best you can and have supporting information to defend the choices you make.

The greatest challenge of structured data sources is that they are often very large, containing millions, billions or even trillions-of rows of data corresponding to hundreds of GB or TB of data. Databases of this size are expensive to collect and produce, and they also require enormous expertise to analyze and clean up. Involving qualified eDiscovery experts specializing in large-scale analysis and search of structured and unstructured data is critically important as is ensuring the efficient use of their time.

A final challenge is that structured data sources frequently contain sensitive data. As is true with documents, the court may be willing to make exceptions on producing competitive or proprietary data, but only if you can clearly explain and defend your need. In some cases, opposing counsel will insist on hiring a neutral third party to analyze sensitive data, thus mitigating potential conflicts of interest while still ensuring the collection includes all relevant data.

WHAT SHOULD I DO WITH THE STRUCTURED DATA I COLLECT?

Once your initial collection of structured and unstructured data is complete, find a tool that can search and cross-correlate equally well against both the structured and unstructured data. This may be a tall order, so engage an experienced data analyst to help you with identifying the appropriate cross-analytic, dual-purpose tool. The reality is that there are very few tools that can handle both data types, still fewer that can scale to the volumes demanded by major enterprises For example, suppose you are litigating a medical claims case in which a hospital is suing an insurance provider. While you believe you've collected comprehensive data on all the claims and payments in question, further scrutiny by a data analyst might uncover payments that don't specify the individual claims they covered. As another example, those claims may have been paid or may not have been paid; you'll have to collect more information before you can be certain. Or, perhaps there is a contract or agreement that allows such payment or non-payment).You will need to do a thorough cross-analysis of database records, contracts, and emails to ensure that your assumptions and search criteria are correct. Cases involving complex data often require counsel to spend time on a back-and-forth analysis before enough facts are available to answer the questions involved in the case. The time spent on detailed analysis is well justified, however, since it helps shape the most accurate production possible.

When it comes to production, the meet and confer is generally the best time to negotiate a format agreeable to both sides. The best — and typically the only- practice for large-scale mixed data is to export the raw data into a comma-delimited file. In addition to production format, also consider negotiating a clawback agreement in case sensitive or proprietary data is inadvertently produced. You need to produce both structured and unstructured data in meaningful and correlated format.

Whatever format you choose, make sure to conduct a final review of the collection prior to production. This last appraisal should include both experts who understand the native systems from which the data originated and experts who can analyze the data in the context of the individual case. As with the initial back-and-forth analysis, the resources you spend on final review will be well justified by the comprehensiveness of your collection.

CASE STUDY

Verizon v. OnlineNIC, Case No. 08-02832 (N.D. Cal.)

OnlineNIC, Inc. is one of the largest ICANN- accredited registrars of Internet domain names. In 2008, it was sued by giants Verizon, Microsoft and Yahoo for allegedly registering and using thousands of domains that infringed those plaintiffs' trademarks. The electronic data to be reviewed in the case was massive — it included over 1 million domain names, tens of millions of transactional records (such as registrant data and financial transitional data) and over a million emails. OnlineNIC needed an eDiscovery product that could quickly ingest and cross-correlate both the structured and unstructured data.

OnlineNIC's Defense

OnlineNIC's defense was simple and straightforward — the infringing domains were registered by OnlineNIC's customers — who were individual registrants that were not affiliated with OnlineNIC in any way. Consequently, OnlineNIC maintained that it should not be held liable for any tortious committed by its customers.

The Challenge of Correlating Data

In order to execute its defense, OnlineNIC had to show that the customers who had registered the domains were real persons — not just aliases for OnlineNIC, as alleged by plaintiffs. The only way to do this was to correlate large numbers of financial records (credit card payments) and emails with the data associated with a particular infringing domain. For instance, an account named "lenawoo" - was alleged by plaintiffs to be a mere alias for OnlineNIC. To prove that "lenawoo" i.e. account #232218 was not just an alias for OnlineNIC, it was necessary to show that there were emails, faxes and payment transactions for account #232218 that were traceable to non-OnlineNIC.

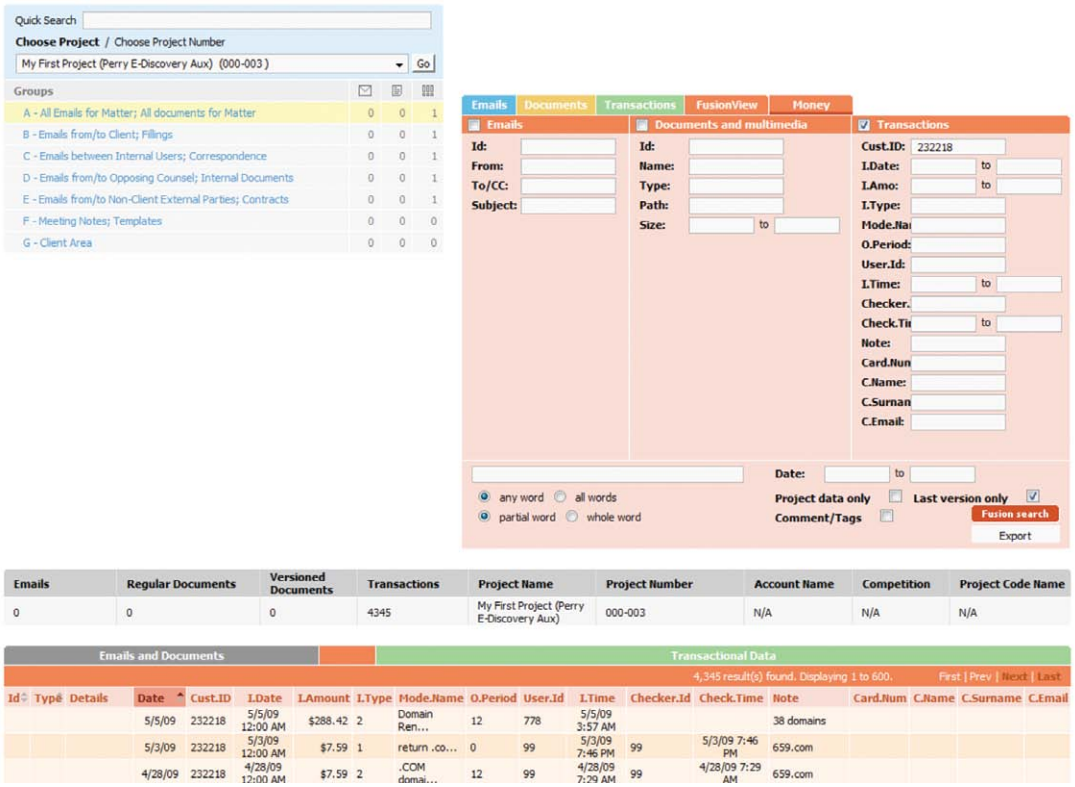

Figure 1. Shows how the LexFusion / BMMsoft EDMT GUI presents all types of data — including Email, Documents, and different types of transactional data. Figure 1, which was produced in litigation, shows over 4,000 payment transactions to a third party - supporting the proposition that 232218 was not an alias for OnlineNIC.

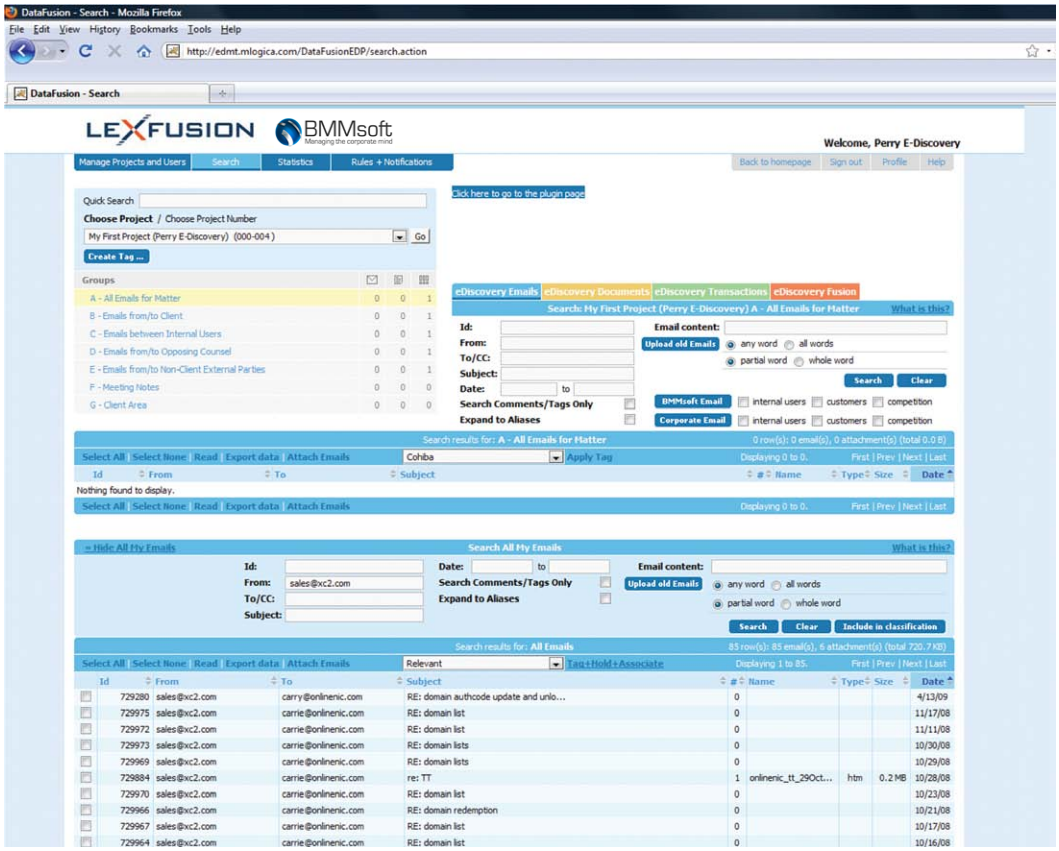

Figure 2. Also produced in litigation, shows an email address (which was listed as a contact email for account 232218) — for which there were over 80 communications back from that email address and OnlineNIC — again, supporting the conclusion that 232218 was not an alias of OnlineNIC. LexFusion / BMMsoft EDMT also enables users to TAG records during review with customizable tags.

Resolution

OnlineNIC was noted in default on the original Verizon complaint. However, with the help of LexFusion / BMMsoft EDMT, the Verizon and Yahoo cases settled amicably, and Microsoft dismissed its case.

For information on our comprehensive Consulting and Education Services to support your Sybase technology initiatives, visit us at www.sybase.com/consulting.

Sybase, Inc.
Worldwide Headquarters
One Sybase Drive
Dublin, CA 94568-7902
U.S.A
1 800 8 sybase

www.sybase.com

SYBASE®
An SAP Company