

Strange Bedfellows ?

Keyword and Conceptual Search Unite to
Make Sense of Relevant ESI in Electronic Discovery*

By Ian Black and Deborah Baron

Introduction

In the brief history of electronic discovery, the latter part of the twentieth century witnessed the demise of paper by a digital hero that emancipated the content of paper documents with OCR and TIFF. This technology added a third dimension to the realm of 2D paper document review and production that lead to a sea change in discovery methods. By many accounts what lies ahead is a three-stage evolution from paper to digital to clustering in order to overcome the problems of volume and complexity of electronically stored information (ESI). The intent of this position paper is to describe the development of the digital hero and methodology that is emancipating the content and context of ESI – conceptual search that spans file formats, languages and technique, and includes keyword search on a common, shared index.

Conceptual Search Theory

'Clustering' is a mathematical breakthrough (even though it's 300+ years old) of which conceptual search is just one advantage discussed in this paper. Conceptual search that offers a wide range of operations on a common scalable infrastructure adds essential dimensionality to legal search in discovery. For example, it delivers contextual understanding of small or massive volumes of ESI and supports unique attorney interests such as early detection of all forms of relevant documents, including those missed by keyword search (an unfortunate but widely recognized issue).¹

Lawyers and investigators need a comprehensive tool to identify these files which are often buried in data sources and collections. A versatile

Early detection of all forms of ESI has become critical as presence of audio, image, video and foreign language files in discovery has grown markedly.

* This paper was presented at the Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings (DESI), June 25, 2008, London, England. http://www.cs.ucl.ac.uk/staff/S.Attfield/desi/DESI_II_agenda.html



conceptual search platform will index these files in the same infrastructure for easy retrieval through a single interface.

How conceptual search tools achieve these results vary widely depending upon their inner workings and underlying theory. Autonomy's approach is based on a unique combination of technologies with theoretical underpinnings that can be traced to Bayesian Inference and Claude Shannon's Principle of Information. Bayes Theorem has become a central tenet of modern statistical probability modeling. Autonomy's advanced pattern-matching technology exploits high-performance probabilistic modeling techniques and extract a document's digital essence to determine the characteristics that give the text meaning. As this technology is based on probabilistic modeling, it does not use any form of language dependent parsing or dictionaries. Words are treated as abstract symbols of meaning and the engine derives its understanding through the context of their occurrence rather than a rigid definition of the language grammar.

Autonomy's approach to concept modeling relies on Shannon's theory that the less frequently a unit of communication occurs, the more information it conveys. Therefore, ideas, which are rarer within the context of a communication, tend to be more indicative of its meaning. It is this theory that enables Autonomy's software to determine the most important (or informative) concepts within a document. Autonomy has extended the theoretical underpinnings with over 100 patents to analyze 1000 content formats and broaden functionality available to users.²

Shortcomings with Keyword Search Alone

As has been seen repeatedly now, overlooking critical content in ESI due to inadequate search technology creates significant risk including court ordered sanctions under FRCP and an attorney's worst nightmare - inadvertent production of privileged documents. In a recent federal civil case, *Victor Stanley, Inc. v. CreativePipe, Inc.*, defendants claimed inadvertent production of privileged documents based on what they argued was a privilege review of text-searchable documents based on an extensive keyword search and a manual privilege review of non-text documents. Defendants also claimed an added burden of too much data to review in the time allotted. Plaintiffs claimed the privilege review was faulty.

US Judge Magistrate Paul Grimm wrote in his opinion, "all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search or relying exclusively on such searches for privilege review." He determined that the defendant waived privilege in large part due to "faulty privilege review of the text-searchable files and by failing to detect the presence of the 165 documents"³ in the production.

"..all keyword searches are not created equal.."

At the root of this issue is the underlying search technology. In his opinion quoted above Judge Grimm aptly points

out that "all keyword searches are not created equal." Many search engines available today miss relevant information because of their performance enhancing shortcuts that are designed to improve the response time and relevancy of information access requests from employees. These shortcuts include 'jump out' which misses potentially relevant documents as it stops looking across an index for potentially relevant information once it estimates a document is unlikely to make the top of section of the results list.

Another shortcut worth mentioning is partial indexing. This is a technique whereby a technology chooses not to index the entire content of the document, but only the first X pages based on assumptions. For example, if a document contains 500 pages of information, the search engine may only index the first five pages. If information relevant to the case appears first on page 6, it will not have been indexed and the search engine may miss this document and others. When these shortcut techniques are applied over even a modest number of files the result is an arbitrary and incomplete set of documents. In legal cases, where a single document has the potential to drastically change the direction of a case, the consequences of these search techniques can be disastrous.

Methodology Does Matter

Judge Grimm advises taking great care in selecting search and information retrieval methodology that is up to the task because failing to do so can be disastrous. He writes in the *VSI v CreativePipe* case that, "The message to be taken from *O'Keefe, Equity Analytics*, and this opinion is that when parties decide to use a particular ESI search and retrieval methodology, they need to be aware of literature describing the strengths and weaknesses of various methodologies, such as *The Sedona Conference Best Practices*, supra, n. 9, and select the one that they believe is most appropriate for its intended task."⁴

The bench is not shy about ordering sanctions under FRCP when parties fail to produce ESI and take action regarding search considerations. In 2007 US Magistrate Judge John Facciola required the parties in the *Disabilities Rights Counsel of Greater WDC v. WDC MTA* case to meet and confer and present him with an agreed search protocol for ESI. In his written opinion the Judge pointed them to "recent scholarship that argues that concept searching, as opposed to keyword searching, is more efficient and more likely to produce the most comprehensive results."⁵

Courts have taken notice of the wide range of ESI sources and data formats in modern organizations and expect a reasonable

"...to conduct a thorough and complete search of both it's hard copy and electronic files in good a good faith effort..."

of both its hard copy and electronic files in a good faith effort to uncover all responsive information in its 'possession custody or control.'⁶

Outside counsel should be on the alert for their client's failures to uncover responsive data as it creates risk of FRCP sanctions for their firm as well. In a current federal civil case, *R & R Sails v Ins. Co. of Pa.*, plaintiff argued that negligent failure by the defendant to locate and produce a claim log responsive to the plaintiff's discovery request was cause for sanctions. The US Magistrate Judge Louisa S. Porter agreed. In this case, defendant and counsel had certified that discovery was complete and no claims log existed.

However the defendant later found a claims log database, and then subsequently found claim log entries on the defendants PC. Counsel turned over a report generated from the database to plaintiffs that was later found to be incomplete. As certification had already been made and other factors were at play, the judge ordered monetary sanctions and recommended non-monetary sanctions for client and counsel under Rule 26 based on inadequate search of and untimely production of ESI.⁷

Conceptual Search Advantages to eDiscovery

The amended FRCP Rule 26(a) states "demand an exhaustive search for and identification of sources of discoverable electronically stored information, regardless of form, including email and voice content for disclosure." ⁸ Voice recordings are a growing form of critical digital evidence from call centers in consumer products liability cases to call recordings in regulated industries. For example, in a dispute between two large banks the defendants "failure to retain audio recordings of its traders' telephone calls was sanctionable". In the judges opinion the "appropriate sanction was adverse inference jury instruction."⁹ Email and voice communications files are more critical and complex than ever before, and legal technology consumers require scalability and analytical tools to more effectively understand and manage them.

A conceptual search platform with built-in analytics enables a comprehensive and efficient discovery process. Analytics can enable early detection of key custodians and ESI regardless of language and form, and rapid culling and pre-review of key custodian data for early case assessment (ECA). Using advanced analytics in the early stage of eDiscovery is invaluable. It assists lawyers and investigators

and defensible practice for a comprehensive search across them. In Judge Facciola's memorandum opinion in the O'Keefe case he cites the court order requiring "the government to conduct a thorough and complete search

to expose communications links and uncover hidden custodians and gaps in email traffic.

A truly useful tool will display these communications patterns in a graphical form that doesn't require a PHD to understand. Mere mortals should be able to view and quickly make sense of the visualizations to better assess risk and more efficiently review documents. The result is a reasonable and defensible search and discovery methodology to pinpoint documents that support your case, rapidly filter non-responsive items and reduce the risk of failures to comply with the FRCP.

If your business is outside the US should you be concerned with FRCP? The answer is "Yes" if your organization conducts business in the US and/or has operations in the country, and you are involved in a legal dispute or US government investigation. If nothing else, think of the watershed *Zubelake v UBS* rulings, which predate the amended FRCP but contributed to their formulation.

If your organization is based outside of the US entirely should you be concerned about keyword search and this new methodology? Again, the answer is a simple yes. Keyword search does not provide you the results needed to effectively pinpoint the ESI that will assist you to defend your case. You will be burdened with over-inclusiveness as well as under-inclusiveness. Additionally, cross border disputes with multilingual ESI and data privacy issues are perhaps the most common reason for larger organizations to adopt a versatile conceptual search platform that supports language independence and early detection of private data.

How important is a multi-dimensional approach to search at the front end of a legal matter or investigation? Once the duty to preserve attaches, the race is on to preserve and ultimately collect in a manner that is FRCP compliant. The biggest issues are over-collection, cost, and spoliated data due to failures to preserve. The same is true for proactive custodian information management, a practice some organizations follow for serial custodians or regulatory reasons, actively archiving their email and files stores in real time.

Relying on keyword search alone in the early phases of a legal or regulatory matter presents the very same limitations as those described above. In addition, it creates data privacy issues for organizations outside of the US because it lacks the ability to distinguish context and usage of the keywords in a document.

The Best of Both Worlds: A Combined Approach

Utilizing flexible and adaptive conceptual technology at the time of preservation to narrow the volume of ESI from custodians allows organizations to...reduce the cost of eDiscovery

of data beginning at its source, and preserve ESI-in-place to allow for reduction in scope before collection. Imagine the opportunity to reduce ESI volume by applying holistic, robust techniques during preservation and collection on custodian desktops, laptops and file shares, in a defensible manner to greatly mitigate over collection and lower costs!

On the other end of the eDiscovery spectrum is production and quality control. Conceptual search techniques along with keyword and Boolean are used in a clever manner to check for privileged and confidential information in a production set. Sample docs are understood conceptually and are used to locate others “like this” with the intent of avoiding inadvertent production of privileged and confidential information. In the case *VSI v Creative Pipe*, Judge Grimm points out that the defendant did not conduct a quality check on their production before releasing it. In his opinion the defendant “failed to demonstrate that there was quality-assurance testing” of their production documents.¹⁰

The combination of keyword and conceptual search and retrieval in a single tool based on a shared index and infrastructure has the unique benefits of speed, scale and extesibility.

Utilizing flexible and adaptive conceptual technology at the time of preservation to narrow the volume of ESI from custodians allows organizations to gain efficiency, mitigate the risk of spoliation and reduce the cost of eDiscovery. Using this approach, an organization will have the machinery it needs to reduce the volume

Equally important to making sense of ESI in discovery are essential human factors such as ease of use via familiar keyword entry. This is a reasonable and defensible technology assisted methodology that reduces risk of FRCP sanctions and inadvertent production of privileged information to an adversary.

The examples discussed in this paper are made in the spirit of progress and an attempt to illustrate an urgency to bring forward more versatile conceptual search methodology and technique to better assist litigators and investigators in making sense of complex and voluminous ESI. This approach, unlike paper review which will one day become extinct, will drive keyword and conceptual search techniques to fuse together and enhance lawyers and their clients ability “to efficiently and efficaciously conduct searches for relevant documents in heterogeneous haystacks of electronic data.”¹¹ ◀

End Notes

1 The limitations, shortfalls and over inclusiveness of keyword and Boolean search have been documented in numerous papers. See **Paul, George L. and J.R. Baron**, “*Information Inflation: Can The Legal System Cope?*,” 22-24, Richmond Journal of Law and Technology (2006), <http://law.richmond.edu/jolt/v13i2/article10.pdf>.

“A Boolean search is an exact-match engine in that a Boolean search engine will only return documents that exactly match the query, and the documents will be returned in no particular order...If AND is used, then the engine will retrieve only documents which contain every term so joined. Such queries generally return too little. If OR is used, then the search engine will return any and every document which contains any one or more of the so joined terms. Such queries generally return too much...”

“In short, language is a “form of life.” Others have catalogued types of indeterminacy arising from this truth. Thus, it is not surprising that lawyers and those to whom they delegate search tasks may not be particularly good at ferreting out responsive information through the use of simple keyword search terms. Furthermore, people make up words the fly, including new codes that function as language. People in different parts of the country, in different parts of an organization, or in different age groups devise their own private languages for the context of their then current environment. For example, what does POS mean? What is 1337?”

2 The approach Autonomy takes is that of format agnosticism that enables organizations to benefit from automation without losing manual control. This complementary approach allows automatic processing to be combined with a variety of human controllable overrides. The technology is a complete scalable, modular software infrastructure that forms an understanding of the actual content of any type of information, text or voice-based, structured or unstructured, regardless of where it is stored, the format it has been created with or the applications associated with the data. This is why the technology provides “Integration Through Understanding”,

By aggregating more than 1000 content formats from 400 enterprise resources Autonomy allows organizations to make sense of information from the widest range of sources, including unstructured content like HTML pages, Office files, email, XML and structured data such as Oracle. The software penetrates the information silos in an organization by offering deep integration into EMC/Documentus, Lotus Notes, Exchange, RDBMS, file servers and more.

3 US Magistrate Judge Paul Grimm, in *Victor Stanley, Inc. v. Creative Pipe, Inc.* defendants claimed inadvertent production of documents following privilege review of text-searchable documents using an extensive keyword search technique and manual privilege review of ‘non-text’ documents. Judge Grimm states “First, the Defendants are regrettably vague in their description of the seventy keywords used for the text-searchable ESI privilege review, how they were developed, how the search was conducted, and what quality controls were employed to assess their reliability and accuracy.” And he continues, “As will be discussed, while it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI, all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search or relying exclusively on such searches for privilege review.” ... “Common sense suggests that even a properly designed and executed keyword search may prove to be over-inclusive or under-inclusive, resulting in the identification of documents as privileged which are not, and non-privileged which, in fact, are.”

Plaintiffs claimed a faulty privilege review and the judge wrote, “Thus, according to the Plaintiff, the Defendants have waived any claim to attorney client privilege or work-product protection for the 165 documents at issue because they failed to take reasonable precautions by performing a faulty privilege review of the text-searchable files and by failing to detect the presence of the 165 documents, which were then given to the Plaintiff as part of Defendants' ESI production. As will be seen, under either the Plaintiff's or Defendants' version of the events, the Defendants have waived any privilege or protected status for the 165 documents in question.”

Victor Stanley, Inc. v. CreativePipe, Inc. D.Md.,2008. ---, 2008 WL 2221841 (D.Md.) May 29, 2008. at 3-4

4 Id. At 6, 5

5 Judge Facciola orders the parties to meet and confer and outline a search protocol for a large volume of ESI. He writes: “how will they be searched to reduce the electronically stored information to information that is potentially relevant? In this context, I bring to the parties' attention recent scholarship that argues that concept searching, as opposed to keyword searching, is more efficient and more likely to produce the most comprehensive results.” See **George L. Paul & Jason R. Baron**, *Information Inflation: Can the Legal System Adapt?* 13 Rich. J.L. & Tech. 10 (2007). *Disability Rights Council of Greater Washington v. Washington Metropolitan Transit Authority*, D.D.C.,2007June 1, 2007, 242 F.R.D. 139, at 10

6 MEMORANDUM OPINION, JOHN M. FACCIOLA, United States Magistrate Judge

By his Order of April 27, 2007, Judge Friedman required the government to conduct a thorough and complete search of both its hard copy and electronic files in "a good faith effort to uncover all responsive information in its 'possession custody or control.'" *United States v. O'Keefe*, No. 06-CR-0249, 2007 WL 1239204, at *3 (D.D.C. April 27, 2007) (quoting Fed.R.Crim.P. 16(a)(1)(E)).

7 *Client and Counsel Jointly and Severally Liable for Monetary Sanctions Based on Inadequate Search for and Untimely Production of ESI; Evidentiary Sanctions Also Recommended* Posted on June 5, 2008 by **K&L Gates** at <http://www.ediscoverylaw.com/2008/06/articles/case-summaries/> and "Plaintiff argues that Defendant's representations to Plaintiff and to the Court that a claim log responsive to Plaintiff's discovery request did not exist, violated Rule 26(g) and represent at least a negligent failure by Defendant to locate, review and produce discovery." As a result, the magistrate judge issued an order that defendant and its counsel were jointly and severally liable for attorneys' fees and costs

B. Sanctions Are Warranted Under *Federal Rule of Civil Procedure 37*

Non-monetary sanctions: Federal Rule of Civil Procedure 37(c) provides remedy for a party's failure to supplement its disclosures under 26(e). Rule 26(e) requires that parties supplement their initial disclosures "in a timely matter if the party learns that in some material respect the disclosure or response is incomplete." "Rule 37(c) instructs courts to disallow use of the information that was withheld and/or order the payment of costs and fees caused by the failure to supplement disclosures."

R & R Sails Inc. v. Ins. Co. of Pa., 2008 WL 2232640 (S.D. Cal. Apr. 18, 2008) at 2-4

8 Rule 26(a), Early Disclosures; "Meet and Confers" and Identification. Federal Rules of Civil Procedure (FRCP) (2006) Amendments to the discovery rules demand an exhaustive search for and identification of sources of discoverable electronically stored information, regardless of form, including email and voice content for disclosure. As a result of the search, a "copy of, or a description by category and location" of all electronically stored information that "the disclosing party may use to support its claims or defenses" must be presented. In the case of email, this disclosure may require references to email that may be stored on backup tapes, employee PCs, and/or Blackberry devices.

9 "Securities lender's counsel failed to conduct reasonable inquiry into existence of recordings of its trader's telephone calls prior to responding to request for recordings in action alleging that lender perpetrated fraudulent securities loan and market manipulation scheme, and thus lender was subject to discovery sanctions."

Background: Intermediate lenders brought actions alleging that securities lenders perpetrated fraudulent securities loan and market manipulation scheme. Plaintiffs moved for sanctions, and defendant moved for attorney fees and costs.

Holdings: The District Court, Kyle, J., adopted report and recommendation of Boylan, United States Magistrate Judge, which held that:

- (1) defendants' duty to preserve relevant information commenced they received order indicating that bankruptcy court was investigating alleged scheme;
- (2) lender's failure to retain audio recordings of its traders' telephone calls was sanctionable;
- (3) appropriate sanction was adverse inference jury instruction; and
- (4) lender's counsel failed to conduct reasonable inquiry into existence of recordings.

*E*TRADE SECURITIES LLC, Plaintiff, v. DEUTSCHE BANK AG, et al., Defendants; Ferris, Baker Watts, Inc., Plaintiff, v. Deutsche Bank Securities Limited, et al., Defendants.*

Nos. 02-3711(RHK/AJB), 02-3682(RHK/AJB). April 18, 2005.

10 *Id.* at 6, "Additionally, the Defendants do not assert that any sampling was done of the text searchable ESI files that were determined not to contain privileged information on the basis of the keyword search to see if the search results were reliable."

11 DESI II Background Paper Feb. 29-2, <http://www.cs.ucl.ac.uk/staff/S.Attfield/desi/index.html>

Autonomy and the Autonomy logo are registered trademarks or trademarks of Autonomy Corporation plc. All other trademarks are the property of their respective owners.

About Autonomy

Autonomy Corporation plc (LSE: AU. or AU.L) is a global leader in infrastructure software for the enterprise and is spearheading the meaning-based computing movement. Autonomy's technology forms a conceptual and contextual understanding of any piece of electronic data including unstructured information, be it text, email, voice or video. Autonomy's software powers the full spectrum of mission-critical enterprise applications including information access technology, BI, CRM, KM, call center solutions, rich media management, information risk management solutions and security applications, and is recognized by industry analysts as the clear leader in enterprise search.

Autonomy's customer base comprises of more than 17,000 global companies and organizations including: 3, ABN AMRO, AOL, BAE Systems, BBC, Bloomberg, Boeing, Citigroup, Coca Cola, Daimler Chrysler, Deutsche Bank, Ericsson, Ford, GlaxoSmithKline, Lloyds TSB, NASA, Nestle, the New York Stock Exchange, Reuters, Shell, T-Mobile, the U.S. Department of Energy, the U.S. Department of Homeland Security and the U.S. Securities and Exchange Commission. Autonomy also has over 350 OEM partners and more than 400 VARs and Integrators, numbering among them leading companies such as BEA, Citrix, EDS, IBM Global Services, Dassault Systemes, Satyam, Sybase, Symantec, TIBCO, Vignette and Wipro. The company has offices worldwide.

The Autonomy Group includes: ZANTAZ, the leader in the archiving, e-Discovery and Proactive Information Risk Management (IRM) markets; Cardiff, a leading provider of Intelligent Document solutions; etalk, award-winning provider of enterprise-class contact center products, Virage, a visionary in rich media management and security and surveillance technology and Meridio, a leading provider of records management software.

Autonomy Inc.

One Market, Spear Tower, 19th Floor,
San Francisco, CA 94105, USA
Tel: +1 415 243 9955
Fax: +1 415 243 9984
Email: info@us.autonomy.com

Autonomy Systems Ltd

Cambridge Business Park,
Cowley Rd, Cambridge CB4 0WZ, UK
Tel: +44 (0) 1223 448 000
Fax: +44 (0) 1223 448 001
Email: autonomy@autonomy.com

Other Offices

Autonomy has additional offices in Antwerp, Barcelona, Beijing, Bogota, Boston, Buenos Aires, Calgary, Cambridge, Chicago, Dallas, Darmstadt, Kuala Lumpur, London, Madrid, Mexico City, Milan, Munich, New York, Oslo, Ottawa, Paris, Pleasanton, Rome, San Francisco, Santa Clara, Shanghai, Singapore, Santiago, Sao Paulo, Stockholm, Sydney, Tokyo, Utrecht and Washington, D.C.

