

DESI II

Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings

Hosted by UCL Interaction Centre
University College London

25th June 2008

Organising Committee

Simon Attfield, University College London
Jason R. Baron, National Archives and Records Administration
Stephen Mason, Institute of Advanced Legal Studies
Douglas W. Oard, University of Maryland

Schedule

08:30 Registration Opens

9:00am – 10:30am

Session 1: The DESI challenge

Chair: Simon Attfield

Welcome - Ann Blandford, UCL Interaction Centre

Keynote - Jeane A Thomas, Crowell & Moring

Discussant - Mark Sanderson, Univ. of Sheffield

E-discovery Research at TREC – Douglas W Oard, Univ. of Maryland

US Caselaw Update - Jason R Baron, US National Archives and Records Administration

New Challenges Require New Solutions & Sedona International Working Group Update - Reza Alexander, DLA Piper, UK

Moderated discussion (goals for the day)

10:30 Coffee

11:00 – 12:30

Session 2: Papers

Chair: Doug Oard

Towards an Expanded Model of Litigation - Jacki O'Neill, XRCE

"One Size Fits All" Software Does Not Fit in the Legal Sector - Kelly KJ Kuchta, Forensics Consulting Solutions, LLC

Strange Bedfellows? Keyword and Conceptual Search Unite to Make Sense of Relevant ESI in Electronic Discovery - Ian Black, Autonomy Group

Term Testing: A Case Study - Chris May, IE Discovery, Inc.

Automated Legal Sensemaking: The Centrality of Relevance and Intentionality – Robert S Bauer, H5

Discussant - Craig Carpenter, Recommind

Moderated discussion

12:30 – 13:45 Lunch

Proposed discussion groups (with moderators)

Language processing (Yunhyong Kim, Univ. of Glasgow)

Information retrieval (Mark Sanderson)

Sensemaking (Ian Ruthven, Univ. of Strathclyde)

Vendors (Reza Alexander)

Barristers (Jeane Thomas)

13:45 – 15:20

Session 3: Papers

Chair: Jason Baron

E-discovery Viewed as Integrated Human-Computer Sensemaking: The Challenge of 'Frames' - Simon Attfield, UCL Interaction Centre

Jigsaw: Investigative Analysis on Text Document Collections through Visualization - Carsten Görg, Georgia Institute of Technology

Reconstructing Financial Statements - Frank Bennett, Jr., Faculty of Law, Nagoya Univ.

Conceptual Search – ESI, Litigation and the Issue of Language - David T Chaplin, Kroll Ontrack

CaseMap Issue linking in UK Civil Proceedings - Chris Dale, The e-Disclosure Information Project

Discussant - Mounia Lalmas, Queen Mary, Univ. of London

Moderated discussion

15:20 Coffee

15:00 – 17:00

Session 4: The Way Ahead

Chair: Doug Oard

Lunch table report panel Yunhyong Kim, Mark Sanderson, Ian Ruthven,
Reza Alexander, Jeane Thomas

Moderated discussion

Wrap up

18:00 No-host reception & dinner at:

Paradiso, 35 Store Street London, WC1E 7BS
020 7255 2554

Contents

Session 2 Papers

- 1 **Towards an Expanded Model of Litigation**
Vitorio Benedetti, Stefania Castellani, Antonietta Grasso, Dave Martin and Jacki O'Neill, *Xerox Research Centre Europe (XRCE)*
- 2 **"One Size Fits All" Software Does Not Fit in the Legal Sector**
Kelly KJ Kuchta, *Forensics Consulting Solutions, LLC*
- 3 **Strange Bedfellows? Keyword and Conceptual Search Unite to Make Sense of Relevant ESI in Electronic Discovery**
Ian Black, *Autonomy Group* and Deborah Baron, *Autonomy ZANTAZ*
- 4 **Term Testing: A Case Study**
Angela Reeves and Chris May, *IE Discovery, Inc.*
- 5 **Automated Legal Sensemaking: The Centrality of Relevance and Intentionality**
Robert S. Bauer, Teresa Jade, Bruce Hedin and Chris Hogan, *H5*

Session 3 Papers

- 6 **E-discovery Viewed as Integrated Human-Computer Sensemaking: The Challenge of 'Frames'**
Simon Attfield and Ann Blandford, *UCL Interaction Centre*
- 7 **Jigsaw: Investigative Analysis on Text Document Collections through Visualization**
Carsten Görg and John Stasko, *Georgia Institute of Technology*
- 8 **Reconstructing Financial Statements**
Frank Bennett, Jr., *Faculty of Law, Nagoya University*
- 9 **Conceptual Search – ESI, Litigation and the Issue of Language**
David T. Chaplin, *Kroll Ontrack*
- 10 **CaseMap Issue Linking in UK Civil Proceedings**
Chris Dale, *The e-Disclosure Information Project*

Supplemental Reading

- 11 **Compilation of Selected Recent U.S. Case Law & Commentary Referencing Search & Information Retrieval Methods**
Jason Baron (Updated as of June 15, 2008)
- 12 **US v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008)**
- 13 **Victor Stanley v. Creative Pipe, 2008 WL 2221841 (D. Md.)**

Biographies

Session 2
Papers

1

Toward an Expanded Model of Litigation

Vitorio Benedetti, Stefania Castellani, Antonietta Grasso, Dave Martin, Jacki O'Neill

Xerox Research Centre Europe
6, chemin de Maupertuis, 38240 Meylan, France
{Firstname.Lastname}@xrce.xerox.com

INTRODUCTION: THE IMPORTANCE OF SOCIO-TECHNICAL DESIGN

The call for contributions for this workshop describes the important new challenges for the legal search community this domain brings. Rather than just understanding the challenges this domain poses in terms of their technical properties, we would like to suggest that understanding these challenges as socio-technical challenges will be important. That is, as well as calling for research on a technical level to address these challenges we are also calling for work to understand the social practices of those involved in e-discovery (ED) and related legal work. A particularly interesting feature of this field is that it is likely that search technologies will (at least semi-)automate responsiveness review in the relatively near term and this will change the way that the work is organised and done in many ways – offering new possibilities for new ways of organising the work. As well as designing those technologies for automating responsiveness review we need to be envisioning how the work will be done in the future, how these technologies will impact the organisation of the case and so on. In this position paper we therefore outline the importance of understanding the wider social context of ED when designing tools and technologies to support and change the work. We would like to reinforce and expand on Conrad's call for IR researchers to understand just what ED entails [2], include the stages that come both before and after core retrieval activities.

The importance of considering the social aspects of work in the design of the technology has been established for some time. Ushering in this 'turn to the social,' and focusing on interface design, Gentner and Grudin [4] described how the GUI has *already* changed from an interface for engineers, representing the engineering model of the machine to one that supported single 'everyman' users (based on ideas from psychology). *From then onwards* the interface has evolved to support groups of users, taking into account the social and organisational contexts of use. This has particular resonance for the design of ED technologies: during ED in particular and the wider legal process there are often many lawyers involved – reviewing documents, determining issues, etc. Even if the way that their work is organised currently is not seen as collaborative in the traditional sense – with individual lawyers working on individual document sets to review them - their work needs to be coordinated and it seems likely that their work could be enhanced by, for example, knowledge of what their colleagues had found, how the case was shaping up, new key terms and facts turned up and so on. Work is often modelled for the purposes of design using process models, but this misses out on the richness and variety actually found when one examines how the work is carried out [3]. Technologies which strictly enforce the process models can often hinder the work, or end up being worked around as was the case with workflow systems since people interpret processes very flexibly to get the work done ([1], [3]). Other studies in other fields have found similar problems when systems are designed on for example cognitive models of how the work is done; they often do not take into account the situated nature of the work and thus they can be very difficult to use [5]. We believe, like [2], that a clear understanding of the social practices of ED is vital for the creation of high-quality, meaningful tools and technologies. We furthermore propose that work practice studies, to be used in combination with other methods, are a central part of getting the detailed understanding of the work practices central to designing useful and intelligent tools. Work practice studies would involve ethnographies, consisting primarily of observation, undertaken of practitioners engaging in the work of ED.

PUTTING THE E-DISCOVERY IN A WIDER CONTEXT

As a first stage of pursuing this research program, we have begun by trying to put the work of ED in a wider context, expanding on existing models like the EDRM one (see Figure 1). In order to propose a new expanded model, which we hope to further validate and refine with work practice studies, we have analyzed the limitation of current models. We take the EDRM model as representative of them, since it was

constructed as part of a common effort of standardization. It seems to us that, while accounting for all the stages, current models of ED mainly focus on the stages up to production. More specifically much of the work of the research community focuses on what is currently the most costly part of the litigation process, e.g. the review for responsive/non responsive documents. For this reason there are many ED tools in support of review. On the other side technology design would benefit from a widening of the current focus to all the *related* case work.

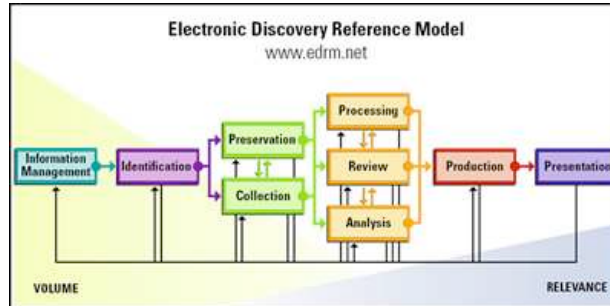


Figure 1: The EDRM model.

We believe this is important for a number of reasons:

- § As new tools change the way the work of ED is done, they could impact on other parts of the process of preparing the case. This should be easier to predict when one keeps in mind the case in it's entirety rather than ED in isolation;
- § tools that support ED might also be relevant in other parts of the process;
- § different parts of the process have implications for how ED can be carried out and as that process changes so there might be consequent changes elsewhere.

In particular there are two areas that need to be better highlighted. The first one concerns the case reasoning activity that ultimately produces the defence or attack line to be used to settle or to go to court. The reasoning about the case is likely to happen from the very start, so we expect it to be important to understand how the review activity informs that aspect in order to design the technology in support of case construction. The second aspect relates to the fact that to our knowledge none of the existing models, including the EDRM, considers the two sides of the process: the plaintiff and the defendant. We believe that this is important when considering design, because technology and reasoning tools could be applied in different ways to determine and find what is and what is not evidence, on the base of these two different perspectives and bodies of knowledge. To conclude, we believe that the design of technology in support of the litigation process is forced to face a number of challenges that are addressed at best by understanding the socio-technical aspects of work through extensive field studies. This research agenda brings additional challenges due to the very confidential and high work pressure nature of these settings, however we believe it is crucial to propose technology that fits with the very sensitive practices of the legal profession.

REFERENCES

- [1] Bowers, J., Button, B., Sharrock, W. (1995): Workflow from within and without. *Proc. ECSCW'05*. 51-66.
- [2] Conrad (2007) E-Discovery Revisited: A Broader perspective for IR Researchers DESI 2007 1st workshop on E-Discovery
- [3] Dourish, P. Holmes, J. MacClean, A. Marqvardsen, P., Zbyslaw, A. (1996): Freeflow: Mediating Between representation and Action in Workflow Systems. *Proc. CSCW'96*. 190-198.
- [4] Gentner, D. & Grudin, J. (1990): Why good engineers (sometimes) create bad interfaces. *Proc. CHI'90*. 277-282.
- [5] Suchman, Lucy (2007): *Human-Machine Reconfigurations: Plans and Situated Actions*. 2nd Ed. Cambridge: Cambridge University Press.

2



“ONE SIZE FITS ALL” SOFTWARE DOES NOT FIT IN THE LEGAL SECTOR

Kelly KJ Kuchta, CEO
Forensics Consulting Solutions, LLC
411 North Central Avenue, Suite 170
Phoenix, AZ 85004

kjkuchta@forensicsconsulting.com

Phone: +1 602-354-2799

Fax: +1 602-992-5292

<http://www.forensicsconsulting.com>

Abstract

It is estimated that by 2011, the amount of electronic data created and stored will grow to 10 times the 180 exabytes that existed in 2006, reflecting a compound annual growth rate of almost 60%*. As the amount of electronically stored data increases and the cost of Electronic Discovery escalate, many companies are rushing to find a “magic pill” that can help them manage records and lower E-Discovery costs in the future. In response to this concern, several software firms have added records management programs to their current software, even Microsoft’s SharePoint is purported to have records management functionality. Unfortunately, discoverable information is taking different forms and our experience suggests that our tried and true methods of identifying responsive data are not effective. These companies claim that with the addition of Records Management they can also help lower the cost of Electronic Discovery required during litigation. Can these “one size fits all” programs actually meet the compliance standards set by the courts? Or, is this “add on” technology making promises that it just cannot deliver on? Can we afford to approach ESI in the context of Electronic Discovery as we have in the past? Is the convergence of Record Management, Compliance, Knowledge Management and Electronic Discovery going to meet in the correct position to meet the legal requirements of ESI? These questions need review.

The Problem

The legal sector currently faces the challenges of the exponential growth of Electronically Stored Information, a corresponding increase in the cost of electronic discovery, and technology is challenging the judiciary that is struggling to define the parameters around electronic discovery.

The electronic discovery market is projected, by the 2006 Socha-Gelbmann Survey, to top \$3B in 2008 which has caught the attention of a number of major technology organizations. For example, Microsoft is including records management and electronic discovery processes in its SharePoint platform. While collaborative systems like this offer great advantages they also add the challenges of identifying exactly who viewed or participated in the modification of documents. Without a full vetting of the capabilities and limitations of these systems



companies could be putting themselves in jeopardy if they depend on them as a source of electronic discovery. In addition, electronic discovery can also include text messaging, voice mail, copier memory, PDA / Blackberry storage, memory sticks, and any historical enterprise data. The “one size fits all” records management system does not encompass these records and well could prove inadequate to meet the demands of the court.

The Position

The entry of major technology providers in the electronic discovery market may be very good for the industry. However, that will not be true if the primary vehicle is an add-on to a records management program. The industry needs to take a hard look beyond “one-size-fits-all” solutions to those that can truly keep pace with the challenges to be faced.

★ *The Diverse and Exploding Digital Universe – An IDC White Paper March 2008*

3

Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings (DESI)

Strange Bedfellows? Keyword and Conceptual Search Unite to Make Sense of Relevant ESI in Electronic Discovery.

By Ian Black and Deborah Baron

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 we have before us is a three-stage evolution from paper to digital to clustering in order to overcome the problems of volume and complexity of 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.

‘Clustering’ is a mathematical breakthrough (even though it’s 300+ years old) of which conceptual search is just one advantage we are discussing here. 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

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. Lawyers and investigators need a comprehensive tool to identify these files which are often buried in data sources and collections. A versatile 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 informs our vision and 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. We use advanced pattern-matching technology to exploit 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. We have extended the theoretical underpinnings with over 100 patents to analyze 1000 content formats and broaden functionality available to users².

As we've 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.

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.

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 and defensible practice for a comprehensive search across them. In Judge Facciola's memorandum opinion on the O'Keefe case he cites the court order requiring "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.'"⁶

Outside counsel should be on the alert for their client's failures to uncover of 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. And 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 defendant later found a claims log database and later still claim log entries on 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

The amended FRCP Rule 26(a) “*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 and damages in excess of \$600 million. 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 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? 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? 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. And finally 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 there are over-collection, spoliated data due to failures to preserve and cost. 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 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. A good example is searching for the word shred and its extension in an organization's data. (shred example here)

Leveraging 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 of data beginning at its source, and preserve ESI-in-place allowing 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 *VSI v Ca*, 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 extensibility. Equally important to sensemaking of ESI in discovery are essential human factors such as ease of use via familiar keyword entry, 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. In addition this approach, unlike paper review which will one day become extinct, will drive keyword and conceptual search techniques to fuse together and live on to assist lawyers and their clients "to efficiently and efficaciously conduct searches for relevant documents in heterogeneous haystacks of electronic data"¹¹

#####

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 holds over 100 patents and its 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 including eDiscovery 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, Lloyd 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 300 OEM partners and more than 400 VARs and Integrators, numbering among them leading companies such as BEA, Business Objects, Citrix, EDS, IBM Global Services, Novell, Satyam, Sybase, Symantec, TIBCO, Vignette and Wipro. The company has offices worldwide.

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

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., 2007 June 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." At

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

4



Term Testing – a Case Study

IE Discovery, Inc.

1. Purpose and Background

The litigation world has many examples of cases where the volume of Electronically Stored Information (ESI) demands that litigators use automatic means to assist with document identification, classification, and filtering. This case study describes one such process for one case. This case study is not a comprehensive analysis of the entire case, only the Term Testing portion.

Term Testing is an analytical practice of refining match terms by running in-depth analysis on a sampling of documents. The goal of term testing is to reduce the number of false negatives (relevant / privilege document with no match, also known as “misdetections”) and false positives (documents matched but not actually relevant / privilege) as much as possible.

The case was an employment discrimination suit, against a government agency. The collection effort turned up common sources of ESI: hard drives, network shares, CDs and DVDs, and routine e-mail storage and backups. Initial collection, interviews, and reviews had revealed that a few key documents, such as old versions of policies, had not been retained or collected.

Then an unexpected source of information was unearthed: one network administrator had been running an unauthorized “just-in-case” tracer on the email system, outside the agency’s document retention policies, which created dozens of tapes full of millions of encrypted compressed emails, covering more years than the agency’s routine email backups. The agency decided to process and review these tracer e-mails for the missing key documents, even though the overall volume of relevant documents would rise exponentially.

The agency had clear motivation to reduce the volume of documents flowing into relevancy and privilege reviews, but had concerns about the defensibility of using an automated process to determine which documents would never be reviewed. The case litigators and Subject Matter Experts (SMEs) decided to use a process of Term Testing to ensure that automated filtering was both defensible and as accurate as possible.

2. Term Testing Process

The Term Testing process is an iterative approach to refining match terms. A subset of documents is reviewed definitively by SMEs for relevance and privilege, then run through the first pass of match terms to discover false negatives and false positives. Terms are refined and re-run until the results are within

limits of acceptability as defined by the client and the circumstances of the case. The rest of the case study explains the steps in depth, and gives detailed numbers to show affected volumes.

2.1. Identify Document Sample

The complete collection of tracer e-mails is estimated to be approximately 10 million documents (exact numbers are unknown because the e-mails are stored in a compressed format, each compressed file unpacks to a different number of e-mails, and not all tapes were uncompressed before the project was put on hold).

Given that the tracer e-mails were collected by date (versus most collections, which are primarily by custodian or location), date was the best criteria to ensure a representative sample. Uncompressed files were selected from each year in the collection, covering different months in case there was an unexpected seasonal factor. The files from the chosen dates ended up comprising 15,220 documents, approximately 0.15% of the overall collection.

The decision to keep the sample small was practical as well as statistical in nature. For practical purposes, because all the documents in the sample required in-depth human analysis from a small team of experts, a collection of more than about 15,000 could jeopardize the target duration of three weeks for the Term Testing process. Statistically, the attorneys hoped to identify the key documents within the first 1,000,000 documents processed, which gave a sampling of approximately 1.5% of the overall collection, a much more common sample size.

The process to identify the documents only took one day.

2.2. SME Review

Two SMEs reviewed each document for relevance and privilege, in a double-blind review so neither SME knew how the other had marked the document. Documents with conflicting markings were then reviewed by a panel of SMEs charged with resolving conflicts. Fortunately, only 326 documents were in dispute (2.1% of the sample), so the resolution process took only one day. These SME review decisions became the standard of correctness for the rest of the process, so the results of the automated term matching were compared against the SME review decisions to determine false negatives and false positives.

2.3. Search Terms List 1

The SME team created a first draft of search terms, known as "List 1". Terms were identified for both relevance and privilege simultaneously, since the match technology only needed to be run once to determine hits on both sets of terms. One SME made a first draft of the term list, and other team members added to the list. For the first pass, all ideas were included.

Then technical staff translated the English-version of the term to a technical, regular-expression format of the term. The regular expression technology allows for one term to cover multiple spellings where

required. For example, the name “Stephen” is a regular expression of “Ste(v|ph)e(n)?” which catches any of the following versions of the name: Steve, Steven, Stephe, Steven.

The creation of List 1 happened before the other steps in the Term Testing process began, so did not add any days to the process.

2.4. List 1 Analysis and List 2 creation

List 1 was run against all the sample documents for search term hits. The SME team then analyzed both false negatives (relevant / privilege document with no match) and false positives (documents matched but not actually relevant / privilege). The agency had already decided that the risk of omitting potentially relevant documents outweighed the costs of reviewing a higher volume of documents, so the primary focus was on eliminating false negatives down to zero.

The group was pleasantly surprised to find that the first pass only resulted in 921 documents which were false negatives (6.1%), and 1,892 documents which were false positives (12.4%). The SME group distributed the false matches and combed through the documents to find new terms, and to analyze terms as candidates to be removed.

The technical staff analyzed the term hits to see if any terms were unnecessary, which would have been true one of three ways: either (a) the term did not match any documents in the sample, or (b) the term only matched documents where other terms also matches, so could be unnecessary, or (c) the term produced so many false positives that it was not helpful to include. No terms fell into category (a). The few terms which fell into category (b) were deemed too necessary to discard. Only 4 terms were considered in category (c), but only 2 terms were actually deemed unnecessary.

The total change between List 1 and List 2 as the addition of 12 terms, and the elimination of 2 terms, for a total net increase of 10 terms.

This analysis phase of the process took longer than any other portion; it only took a couple of hours to run the terms against the documents, but the analysis of results and creation of List 2 took about five business days to complete.

2.5. List 2 Analysis and List 3 creation

List 2 was run against all the sample documents. The group was very pleased that no documents were found to be false negatives (0.0%), but somewhat discouraged that the false positive rate rose to 4,120 documents (27.1%). The team repeated the same analysis performed after List 1 ran, and agreed on removing 5 of the new terms from the list.

This second analysis only took 3 business days to complete.

2.6. List 3 Analysis

List 3 was run against all sample documents. The result of 0 documents with false negatives (0.0%) was achieved, so even though the false positives were 2,133 documents (14.0%), List 3 was finalized as the list of record to start processing the entire collection.

This third analysis only took 2 business days to complete.

2.7. Ongoing Evaluation

On an ongoing basis, two SMEs dedicated one day per month to review non-relevant documents. Documents were identified for the non-relevant review by randomly selecting approximately 1.0% of the documents processed during the previous month with no relevancy match terms. The goal was to ensure that additional terms were not needed. In the three months where these reviews occurred, no false negative documents were identified, so no term changes were made.

Had new terms been identified, the new terms would have been run against all files, including files which had been analyzed with List 3 terms.

3. Outcome and Summary

As often happens in complex litigation, the case changed mid-stream, and in this case, because case strategies paid off. The tracer e-mail effort started by focusing on a specific 6-month time window most likely to uncover missing key documents. As hoped, some of the missing key documents were found quickly, and produced to opposing council immediately. The revelations from those key documents changed the nature of the matter so fundamentally that the litigators decided to suspend further tracer e-mail efforts indefinitely to focus resources elsewhere. Although the case has not yet been fully resolved, the tracer e-mail effort continues to be on hold.

The team considers the Term Testing to have been successful because, in conjunction with the time window strategy, the right key documents were uncovered quickly, the risk of missing key documents was significantly reduced, and the defensibility of the match terms was greatly improved.

5

Automated Legal Sensemaking: The Centrality of Relevance and Intentionality

Robert S. Bauer, Teresa Jade, Bruce Hedin, Chris Hogan
H5
71 Stevenson St.
San Francisco, CA 94105
{rbauer, tjade, bhedin, chogan}@H5.com

1. Introduction

In a perfect world, discovery would ideally be conducted by the senior litigator who is responsible for developing and fully understanding all nuances of their client's legal strategy. Of course today we must deal with the explosion of electronically stored information (ESI) that never is less than tens-of-thousands of documents in small cases and now increasingly involves multi-million-document populations for internal corporate investigations and litigations. Therefore scalable processes and technologies are required as a substitute for the authority's judgment. The approaches taken have typically either substituted large teams of surrogate human reviewers using vastly simplified issue coding reference materials or employed increasingly sophisticated computational resources with little focus on quality metrics to insure retrieval consistent with the legal goal. What is required is a system (people, process, and technology) that replicates and automates the senior litigator's human judgment.

In this paper we utilize 15 years of sensemaking research to establish the minimum acceptable basis for conducting a document review that meets the needs of a legal proceeding. There is no substitute for a rigorous characterization of the explicit and tacit goals of the senior litigator. Once a process has been established for capturing the authority's **relevance** criteria, we argue that literal translation of requirements into technical specifications does not properly account for the activities or states-of-affairs of interest. Having only a data warehouse of written records, it is also necessary to discover the **intentions** of actors involved in textual communications. We present quantitative results for a process and technology approach that automates effective legal sensemaking.

2. Sensemaking and Relevance

We look to cognitive-task-analysis research to characterize the sensemaking behaviors ("making sense" of it all) of a senior litigator conducting a document review. "Sensemaking" is necessary for any decision-making; however, as today's information environments have become increasingly complex, decision-making has become much more difficult and time-consuming. So understanding massive and diverse content is not just a simple matter of consuming information or finding it faster. More advanced approaches for interacting with information are needed and current keyword-search and data-mining methods simply cannot meet these needs. [1] Indeed, the courts have recently ruled that the drastic oversimplification of the e-Discovery task as a simple search exercise is wholly inadequate:

"Whether search terms or 'keywords' will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics ... Given this complexity, for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread." [2]

“Use of search and information retrieval (IR) methodology, for the purpose of identifying and withholding privileged or work-product protected information from production, requires the utmost care in selecting methodology that is appropriate for the task because the consequence of failing to do so, as in this case, may be the disclosure of privileged/protected information to an adverse party, resulting in a determination by the court that the privilege/protection has been waived.” [3]

Indeed, the effective modeling of the e-Discovery task requires ‘making sense’ of the salient aspects of the senior litigator’s sensemaking efforts. This task description holds whether a computer system is used or a team of human surrogate reviews conducts the review (with or without technology support.) Therefore, the proper framework for considering how to tag relevant documents within a discoverable document population must incorporate two iterative sensemaking loops, shown schematically in Figure 1.

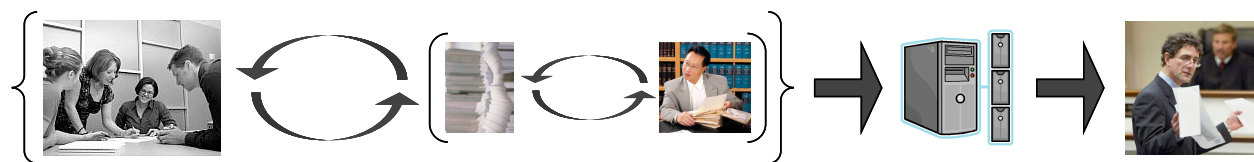


FIGURE 1: Two sensemaking efforts are involved in providing a legal team with an effective document set from a large document population of electronically stored information (ESI:) ‘Making sense’ {outer braces} of the sensemaking activities of the senior litigator [inner brackets.] Machine automation of e-Discovery is shown as the computation over the ESI utilizing the knowledge-based system (KBS) developed by an interdisciplinary team.

The primary sensemaking loop of e-Discovery involves the senior litigator who inherently evaluates document relevance by assessing the intentions of the document’s author relative to the legal strategy of the case [shown in Fig. 1 within the inner brackets.] Any effort that scales beyond a single litigator must simulate this sensemaking activity as closely as possible. Consistent with many other findings [4,] any method that depends primarily on human review fails to transfer properly the requisite knowledge of the senior litigator’s sensemaking into a consistent, reproducible document review. We overcome this inherently human limitation by utilizing a multi-disciplinary team of linguists, lawyers, and subject matter experts to codify their meta-sensemaking model {shown in Fig. 1 within the outer braces,} into a knowledge-based system that replicates the litigator’s primary sensemaking [inner brackets]. Rigorous relevance criteria and in-process measurements of statistically-valid ESI samples are required to assess and ensure accuracy. The document set needed by the legal team can then be produced without further human participation by applying the KBS to the entire ESI document population; this automation often employs a massively distributed computational infrastructure because the ESI scale is typically massive.bb

As Russell, et. al. describe: “Sensemaking is simple—it’s the way people go about their process of collecting, organizing and creating representations of complex information sets, all centered around some problem they need to understand.” [5] Clearly, if the “problem” requiring “understanding” is not fully characterized, then the resulting document review will fail. As Russell, et. al put it in their seminal 1993 paper on “making sense” of large, heterogeneous, and often unstructured document content populations: This is a “general phenomena in which part of the job of sensemaking is to establish the goals of the task.” [6] As noted above, the e-Discovery goal must be established by the senior litigator whose authoritative judgment serves as the only true criteria for a successful review.

There are any numbers of ways to establish what we call ‘Relevance Criteria’ (RC) to guide document review. Engineering frameworks abound for capturing requirements; for example, the Institute of Electrical and Electronics Engineers (IEEE) defines a requirement as “a condition or capability needed by a user to solve a problem or achieve an objective.” While this may seem obvious, it is well established that requirements engineering is the hardest single part of building a software system. [7] In recognition of the central role of authoritatively establishing relevance for document review, the 2008 TREC Legal Track has instituted an interactive task that incorporates a ‘Topic Authority’ to “represent the senior litigator who engages the services of an e-discovery firm.” [8] This is the only way to ensure that legal sensemaking is relevant to the proceedings.

3. Sensemaking and Intentionality

The means of establishing the e-Discovery goal (i.e., RC development) must be coupled with an understanding of what the senior litigator would consider to be important evidence that meets the objectives. In addressing this aspect of sensemaking, it is critically important to recognize that the reasoning of senior and junior members of a litigation team is usually quite different. A classical method for understanding the structure of a cognitive activity is to study occupational experts of that activity; for the case of sensemaking, cognitive task research of intelligence analysts has been conducted for at least the past 10 years. Takayama and Card report that senior and junior analyst behaviors are nearly the opposite of each other: “A general trend of more top-down behavior in seniors and more bottom-up behaviors in juniors has become apparent. Senior analysts begin with their own hypotheses and large personal repositories of information before reaching out to more distant sources to fill in gaps or get updates.” [9] The “hypotheses” for a senior litigator is the context for and intentionality of the author of documents comprising the ESI; the “large personal repositories of information” are accumulated by the senior litigator over his/her years of legal experience. This distinction between senior and junior analyses has been characterized by Pirolli as information processing “driven by bottom-up processes (from data to theory) or top-down (from theory to data).” [10] This is central to efforts to automate the sensemaking task because it dictates that replicating senior litigator sensemaking must be rooted not in ‘data mining’ approaches but in systems that reason from a set of “hypotheses.” As noted for the legal domain, these hypothetical constructs characterize the expected intentions of individuals that are involved in the topics of interest in the case.

The crucial insight based on sensemaking research is that in for e-Discovery, senior litigators are NOT reviewing the literal content of text (i.e., bottom-up), but rather the overarching aspects of the situation and the author’s intent (i.e., top-down.) Figure 2 depicts five essential elements required to characterize the intentions that underlay the relevance goals of an e-Discovery effort. [11]

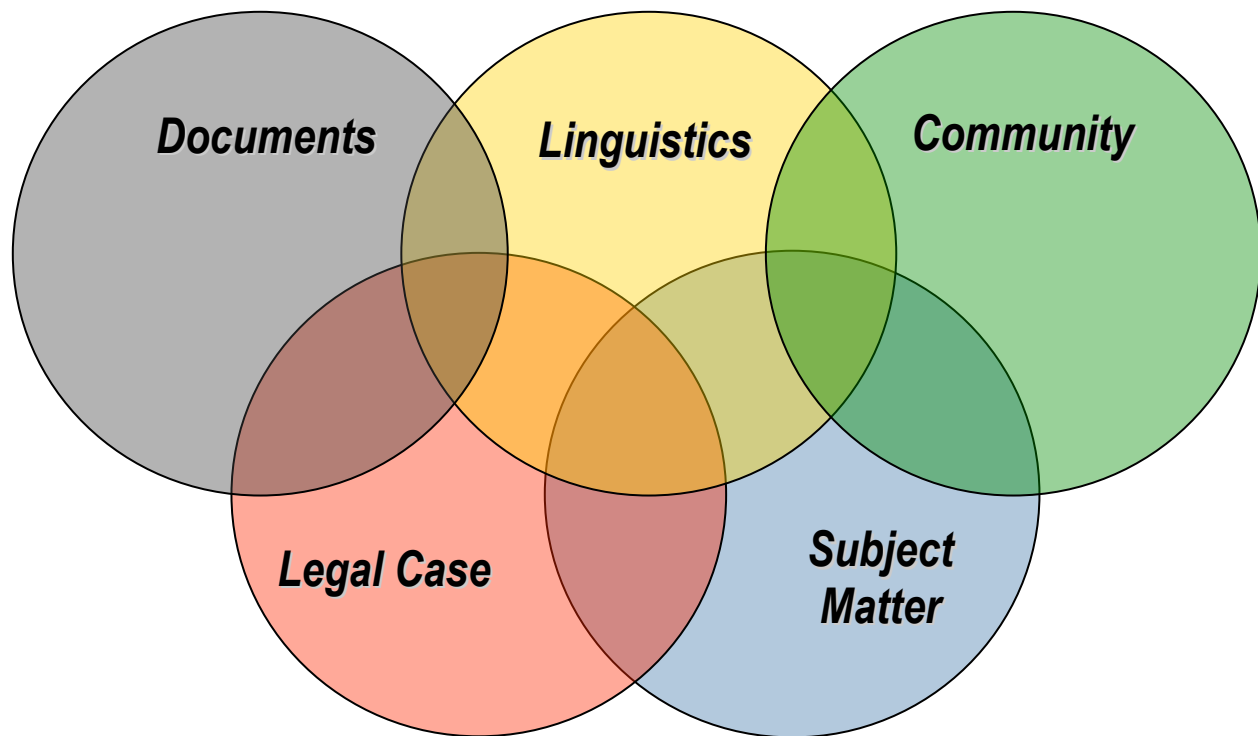


FIGURE 2: Five inter-related aspects that must be considered in characterizing the intention behind the written record. A system or approach that does not at least partially take all of these elements into account is unlikely to consistently achieve the kinds of retrieval results that are acceptable in the legal context.

We find that Relevance Criteria for topics in a review can vary considerably. Some topics are rather simple and readily map to the documents; straight-forward information retrieval techniques can do quite well in finding material that captures such intent. However, other topics are frequently quite complex and require modeling of specific community practices in order to adequately characterize the often subtle, but critical distinctions between relevant and irrelevant documents. Indeed, combining these various dimensions of e-Discovery is a good way to distinguish different approaches and understand their limitations.

a. Documents + Legal Case

In any kind of text-based classification or retrieval system, the documents may simply serve as the target or may also inform the query. In a legal context, the documents contain content and are often associated with metadata (for example, the author of the document, sent date, etc.) The most basic approach to finding relevant documents in the context of litigation involves querying the documents for the topics of interest in the legal case. For example, an attorney might search for “evidence to support a damages claim.” However, the complexity of the legal topics and the fact that the documents were created for other purposes makes a direct mapping of the documents to the legal topics very difficult.

Examples of systems that only consider these 2 dimensions include:

- Manual (human) review conducted by attorneys
- Basic keyword searches targeted to legal issues
- Supervised learning with relevance feedback

b. Documents + Legal Case + Subject Matter

Some legal teams hire subject matter experts to assist them in reviewing particular sets of results. For example, experts in accounting are frequently consulted in cases where calculations regarding damages are required. A key requirement for a

successful system in such a case is a clear and concrete characterization of the target documents. We have found that subject matter can form a bridge from the documents to the legal case. In other words translating the legal case into key subject matter areas creates clarity around which documents will be of interest. For example, when searching for evidence of anti-competitive activities, defining the search in terms of the sales and marketing practices can illuminate where opportunities in the market are unfairly blocked by a competitor.

Examples of systems that consider only these 3 dimensions include:

Subject matter experts review results under legal team direction

Use of Domain-specific lexicons

c. Documents + Legal Case + Subject Matter + Linguistics

Meaning is encoded in language. Linguists are trained to model the salient morphological, syntactic, semantic, sociolinguistic and discourse aspects of the way meaning is encoded in language. The same concepts expressed in different contexts will involve different phraseology. For example, a linguist who analyzes the content of a PowerPoint slide deck can model both the language of that presentation as well as how the presenter might express the same concepts when communicating through email with the boss.

Examples of systems that consider these 4 dimensions include:

Supervised learning (including both relevance feedback and semantic analysis)

Semantic search

d. Documents + Legal Case + Subject Matter + Linguistics + Community

The final key element required to achieve consistently high Recall (R) with high Precision (P) is the characterization of the community in which the documents are created and used. Legal teams often consider various individuals as they prepare for depositions and may compile organization charts and the like to understand roles and responsibilities of key personnel; however, review teams rarely if ever model the processes, states-of-affairs, or idiosyncratic terminology of the communities in which these players participate. In order to understand the document population indicative of the activities of document authors, such rich operating characterizations of communities are vital.

An example system that combines all 5 dimensions is:

Socio-Technical Information Retrieval (STIR) [11]

It is inherent to the top-down sensemaking approach of senior litigators that they review documents with an implicit model of the communities in which actors participate. Individuals create documents in the course of their activities as members of “communities of practice.” [12] We all work in communities over significant period of time wherein specific tools, processes, resources, and language develop; just recall the idiosyncratic acronyms that must be decoded to understand what is transpiring in a new organization you have joined. The language and linguistic forms used to encode meaning are informed by an author’s memberships and roles in their multiplicity of communities. Text cannot be analyzed for the intention of the writer without accounting for the community in which it was created and used.

While the specific framework (such as [13] for 3.d. Communities) may vary, and not every topic requires the same amount of depth, any system that does not take into account all five elements in Figure 2 is unlikely to consistently achieve the kinds of retrieval results that are desirable in the legal context. The nearly universally overlooked sensemaking requirement for e-Discovery that accounts for author intent is the characterization of the community context.

4. Automating Sensemaking: Information Retrieval Processes & Technologies

In order to create an automated legal sensemaking system (shown by computers in Fig.1,) we must consider the meta-sensemaking framework that an expert multi-disciplinary team can use to capture the salient characteristics {shown in braces in Fig.1,} of the senior litigator's sensemaking [shown in inner brackets in Fig.1.] While sensemaking might seem like a vague concept, cognitive task analysis of sensemaking suggests that there is a relatively well-defined structure to the phenomenon. [6, 9] The basic model (called "a learning loop complex") can be summarized in terms of two processes: (1) searching for a representation or framework scheme and (2) actually filling in the framework with the data collected. Attempting to fill in the framework will end up with some data that doesn't fit (called 'residue';) this requires a shift in the representation and then another attempt to fill it in with the data. For our purposes, the knowledge representation used for the multi-disciplinary team's sensemaking is the computational constructs into which the linguistic variations can be expressed and the operations (e.g., Boolean expressions) that enable phrasing of query alternations which capture the meaning being sought in IR computation. The Relevance Criteria developed with the senior litigator (or topic authority in TREC) is the determinant of what constitutes data 'residue' in documents mistakenly tagged as either relevant or irrelevant; this in turn leads to specification of the representational iteration and expressive richness required of the knowledge framework. When the framework accounts for all 5 elements necessary to characterize the intentionality of document authors shown in Figure 2, the information retrieval result is dramatically better than achieved by conventional search technologies plus/or by armies of junior (bottom-up) reviewers.

Results are shown in Figure 3 for iterative development of a litigation review utilizing a hybrid, automated e-Discovery approach addressing all 5 dimensions in Figure 2. Called 'Socio-Technical Information Retrieval' or 'STIR,' [11] it is a knowledge-based system in the classic AI sense of replicating the cognitive sensemaking task of a senior litigator with an automated, computational platform.

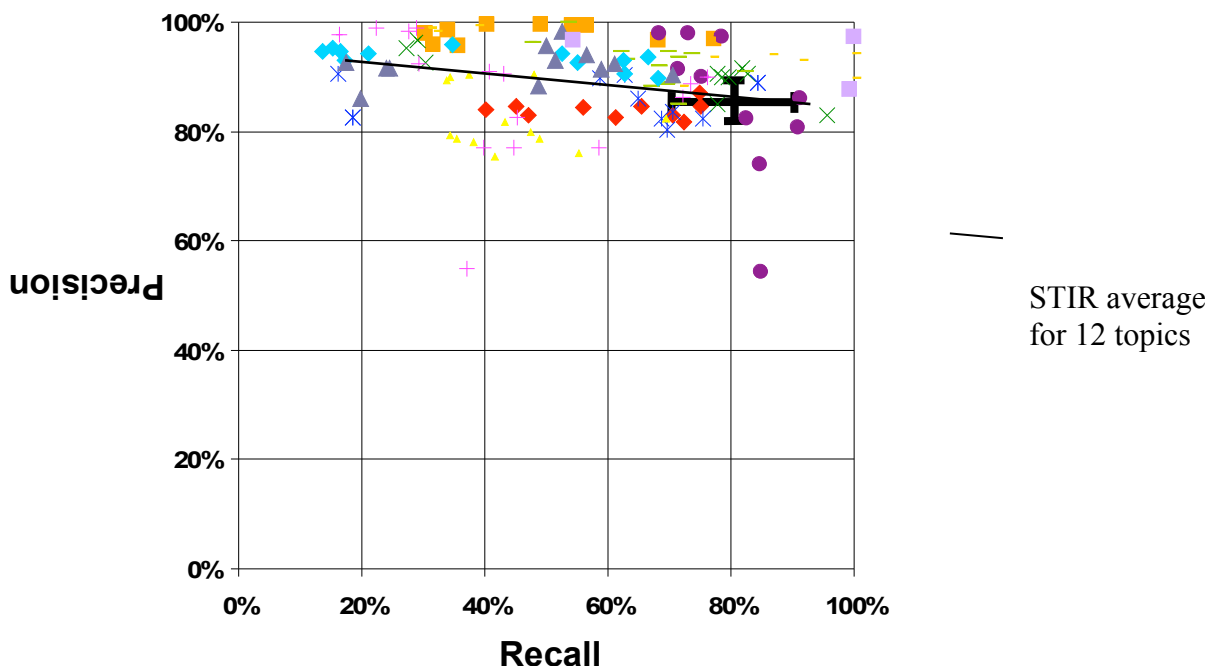


FIGURE 3: Sampled corpus tests for 12 topics during Socio-Technical Information Retrieval (STIR) sensemaking development (i.e., "the process of searching for a representation and encoding data in that representation to answer task-specific

questions.”) Each topic is a different color and all topics move through time from left to right. The black line (from P~93% @R~18% to P~85%@R~80%) is an average for all the topics measured during the building of the knowledge-based system for the legal case.

When STIR development on a case begins, confidence intervals for statistically-valid Recall estimates are rather large, so particular attention is paid to the measure of lower bound. Recall estimation is usually the hardest part of this sort of evaluation; we estimate Recall by comparing the number of documents machine-model tagged as ‘relevant’ to the total number of documents determined to be ‘relevant’ in a double-vetted, quality-controlled, human review of the same, randomly-selected sample population. Over time, retrieval accuracy increases (e.g., increased F-measure [14]) with concomitant narrowing of the confidence intervals (typically +/-6% Recall and +/-3% Precision in our studies.) As would be expected, Precision generally decreases as Recall increases. At the conclusion of the sensemaking development effort utilizing STIR (result shown with confidence intervals of +/-6% R, +/-3% P,) the average quality of the sampled result set for 12 distinct legal issues is dramatically higher than typical TREC interactive task results [15.] The automated STIR sensemaking approach at 80% Recall (i.e., where 4 out of 5 documents of interest to a legal matter would be in the retrieved set) correctly identifies as ‘relevant’ 4 out of 5 documents in the result set. By comparison, AT LEAST ½ of the documents relevant to a legal case will normally be missed by ‘standard’ IR approaches when Precision is 80% or greater.

In practice, post-processing of the result sets by junior (bottom-up) reviews would require 4-5 times the resources to complete the e-Discovery effort compared to the automated sensemaking approach in Fig. 1 for the same Precision and Recall. Often the cost/time requirements for achieving acceptable Recall at high Precision with traditional IR review methods is prohibitive; therefore significant simplifying assumptions are made (often unknowingly) in order to reduce the reviewable document population to a manageable size (e.g., further keyword culling.) Such actions dramatically lower Recall at acceptable Precision in most reviews. Meaningful approaches must explicitly deal with fundamental requirements for making sense of vast amounts of information rather than accommodating to the parameters that are easily manipulated by retrieval tools (e.g., keyword specification, Boolean operators, thesauri context/‘semantics’.)

Figure 3 shows results obtained by linguists using an appropriate, quantitative methodology that reproducibly employs processes, measures, representations, and technologies to craft queries that increasingly produce retrieval results with simultaneous high P & high R. This hybrid, multi-dimensional approach for conducting high-quality, automated, legal sensemaking (a) captures linguistic expertise, (b) characterizes particular practice communities of subject matter experts, and (c) employs some of the latest advances in AI, Natural Language Processing, and massively parallel processing. The penalty for ignoring intentionality and not using a rigorous quality-controlled development process with measurable goals is always greater than the upfront investment required to craft a case-specific, knowledge-based, IR system.

5. Advances in Automated Legal Sensemaking

Development of a fully automated system for legal sensemaking is a process of evolving representations, wherein people seek increasingly effective representations to support the review task and then use them to mechanistically process massive populations of ESI with any human review. This process of representational change during sensemaking is inherently complex, involving hypothesis and test. [16] As depicted in Figure 1, iterative human participation is necessary during this rich query development (i.e., case-specific KBS building;) such iterative query development by experts in the written discourse of practice communities consistently produces high quality, automated “Socio-Technical Information Retrieval” systems.

We refer to such a system as ‘Automated’ because no human interaction is needed for successful conduct of the subsequent IR task on the full corpus.

Future advances must account for the fundamental characteristics of legal sensemaking. To conduct an effective and efficient document review, a system must replicate the sensemaking of senior litigators as a top-down, automated process of searching for a representation and encoding data in that representation to answer case-topic-specific questions. Therefore, two necessary aspects of any scalable, e-Discovery process or technology are (1) establishing explicit criteria for senior litigator relevance and (2) multi-dimensional coding for author intentionality. Clearly, drastic oversimplification of the review task as a keyword search exercise is not capable of the rich, nuanced queries required for sensemaking. Execution of the sensemaking approach requires rigorous measurement and statistically valid, in-process quality control. Without numerical results that characterize the degree of achievement for Precision and Recall, any claim of ‘accuracy’ in automating legal sensemaking is unsubstantiated.

References

1. Palo Alto Research Center’s description of sensemaking research [see http://www.parc.com/about/pressroom/features/sensemaking_0606.html]
2. John Facciola, *United States v. O’Keefe*, No. 06-CR-249, 2008 WL 44972, at *8 (D.D.C. Feb. 18, 2008)
3. Paul Grimm, *Stanley v. Creative Pipe*, No. MJG-06-2662 p. 25-26 (D.MD. May 29, 2008)
4. See seminal study and its citations: D.C. Blair and M.E. Maron, “An evaluation of retrieval effectiveness,” *Communications of the ACM*, 28 (1985), 289-299.
5. D. M. Russell, R. Jeffries, L. Irani, “Sensemaking for the Rest of Us,” CHI workshop on Sensemaking (2008.) [see <http://dmrussell.googlepages.com/sensemakingworkshoppapers>]
6. D. M. Russell, M. J. Stefik, P. L. Pirolli, S. K. Card, “The Cost Structure of Sensemaking,” *Proceeding of InterCHI*, ACM. (1993), 269-276.
7. Carnegie Mellon’s Software engineering Institute [see: http://www.sei.cmu.edu/productlines/frame_report/req_eng.htm]
8. National Institute of Standards and Technology (NIST) - TREC 2008 Legal Track: Interactive Task. [see: <http://trec-legal.umiacs.umd.edu/>]
9. L. Takayama, S. K. Card, “Tracing the Microstructure of Sensemaking,” CHI workshop on Sensemaking (2008.)
10. P. L. Pirolli, “The Cognitive Structure of User Sense Making.” 7th International Workshop of the EU Network of Excellence: DELOS on Audio-Visual Content and Information Visualization in Digital Libraries (AVIVDiLib’05,) Cortona, Italy. [see <http://www.parc.com/research/publications/details.php?id=5455>]
11. R. S. Bauer, T. Jade, M. P. Marcus, 11th International Conference on AI and Law (DESI I Workshop), Stanford, CA - June 4, 2007 [see <http://www.umiacs.umd.edu/~oard/desi-11ws/>]
12. E. Wenger, “Communities of Practice: Learning, Meaning, and Identity,” Cambridge: Cambridge University Press (1998.)
13. A. O. Putman, “Communities,” in *Advance in Descriptive Psychology*, Vol. I, K. E. Davis, ed., JAI Press, Greenwich, CT (1981) 195-209.
14. A single, composite measure of information retrieval accuracy: The weighted harmonic mean of Precision and Recall [see: http://en.wikipedia.org/wiki/Information_retrieval]
15. See for example results in S. T. Dumais & N. J. Belkin, “The TREC Interactive Tracks: Putting the User into Search,” Chapter 6 in E. M. Voorhees & D. K. Harman, ed., “[TREC: Experiment and Evaluation in Information Retrieval](#),” MIT Press (2005) 123-152.
16. G. W. Furnas, “Representational Change in Sensemaking,” CHI Workshop on Sensemaking (2008.)

Session 3
Papers

6

E-discovery viewed as integrated human-computer sensemaking: The challenge of ‘frames’

Simon Attfield

UCL Interaction Centre*
s.attfield@cs.ucl.ac.uk

Ann Blandford

UCL Interaction Centre*
a.blandford@cs.ucl.ac.uk

University College London Interaction Centre, MPEB 8th floor, University College London, Malet Place, London, WC1E 7JE, UK

ABSTRACT

In addressing the question of the design of technologies for e-discovery it is essential to recognise that such work takes place through a system in which both people and technology interact as a complex whole. Technology can promote discovery and insight and support human sensemaking in this context, but the question hangs on the extent to which it naturally extends the way legal practitioners think and work. We describe research at UCL which uses this as a starting point for empirical studies to inform the design of supporting technologies. We report aspects of an interview field study with lawyers who worked on a large regulatory investigation. Using data from the study we describe document review and analysis in terms of a sequence of transitions between different kinds of representation. We then focus on one particular transition: the creation of chronology records from documents. We develop the idea that investigators make sense of evidence by the application of conceptual ‘frames’ (Klein et al’s, 2006), but whilst the investigator ‘sees’ the situation in terms of these frames, the system ‘sees’ the situation in terms of documents, textual tokens and metadata. We conclude that design leverage can be obtained through the development of technologies that aggregate content around investigators’ frames. We outline further research to explore this further.

INTRODUCTION

Electronic Data Discovery (EDD, or e-discovery) has been defined as a process (or series of processes) in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case, or as part of a court-ordered or government sanctioned inspection [Conrad, 2007].

The rapid increase in the volume of electronically stored information within modern enterprises has led to a situation in which preparing for and executing e-discovery represents a considerable challenge for modern organizations and legal companies, and it is one that is set to increase. It is likely that we will increasingly see companies falling prey to legislation if they cannot uncover all electronically stored

information (ESI) relevant to a legal or regulatory matter within a specified timeframe [Baron, 2008].

Advances in digital technologies which have brought about this challenge, however, also offer part of the means for addressing it. The e-discovery technology industry is seeing year-on-year increases in turnover. Software revenues in 2006 were estimated at around \$150 million, with further vigorous growth predicted [Socha & Gelbmann 2007]. Technologies attracting particular interest in this arena include media restoration tools, dedicated document management systems, advanced information retrieval systems (such as concept search and information extraction), information visualization and case analysis tools.

In addressing the question of how to design technology for e-discovery, however, it is essential to recognise that e-discovery work takes place through the operation of a system in which both people and technology interact as a complex whole. In this context, the role of technology is to provide tools and resources that can be usefully appropriated by legal professionals, often working in teams, in constructing strategies and processes that address their goals more effectively. Understanding how technologies can offer additional leverage depends on how those technologies impact on and reshape such systems for the better.

In considering technology developments a significant research object is the e-discovery process viewed as a complex worksystem. Such a perspective becomes particularly pertinent where people are required to engage in intense cognitive activities such as information assimilation, theorising and reasoning, as occurs during the review and analysis of large document collections. As in all branches of knowledge work, technology can promote discovery and insight and support human sensemaking, but the question hangs on the extent to which it integrates within and naturally extends the way that legal practitioners think and work.

We argue that the design of systems to support this kind of work needs to be predicated upon an understanding of the cognitive and social aspects of e-discovery in practice. This

mandates a detailed understanding of the task as it unfolds, including associated processes of sensemaking, teamwork, how people currently coordinate different tools and resources to meet their aims, and what barriers and difficulties arise in doing so. In essence, the need is to examine how work is done in order to speculate how it might be done better [Rasmussen et al 1994].

With this in mind, we are conducting research in field and laboratory settings with the aim of better understanding evidence review and analysis in e-discovery in order to support reasoning about the design of supporting technologies. In this paper we provide an example of that work relating to an interview field study we performed with lawyers who worked on a large regulatory investigation.

In analysing the data from this study two complementary perspectives emerged. The first, which we have reported elsewhere [Attfield et al, 2008a], focuses on how the investigation work was structured. This concerns how the investigators made the investigation tractable by decomposing it into multiple, emerging lines of enquiry or 'issues' distributed across a team. Significant issues relate to how the investigation was decomposed along emerging lines of enquiry resulting from ongoing discoveries, and on the challenge of integrating outcomes from multiple investigation threads to form an integrated perspective.

The second perspective is that of process. Complex knowledge work often occurs in stages which form an iterative sequence of transformations between different kinds of intermediate representation [Attfield, 2008b]. These representations can embody task objectives (such as questions), constructed sub-sets of data with particular meaning (such as search results) or recorded findings and interpretations (such as notes, narratives and structured knowledge representations). As a representation is created

or changed, so it provides raw material for further work, creating new representations and so on. In this way resources act as stepping-stones on an iterative path of sensemaking and a key part of that is the information processing that is performed in order to transition from one representation to another.

In this paper we focus on this second perspective. We first describe the process of document review and analysis in overview in terms of a sequence of transitions between different kinds of representational resource. Next we focus in on one transition in detail, describe how it was done, and use this discussion to reflect on alternative technologies that might offer additional leverage.

INVESTIGATION PROCESS

Based on the interviews, we developed a description of the document review process as shown in figure 1, in the form of a 'process-resource' model. In this figure, boxes represent resources and arrows represent transitions between them. For example, given a set of investigation issues and a document universe, keyword searching (t2) resulted in sets of search results. Given a set of search results, initial manual review (t3) produced a set of documents coded as relevant.

These resources changed constantly and were interdependent but only through transformations (i.e. t1 to t7) which were created by the investigators. The transformations were achieved through some form of information processing, whether this be the investigators reviewing a resource and recording the outcome of their thinking, or by their additional use of automated processing, such as information retrieval.

Each transformation, then, has the effect of using one or more resources in order to shape another, with each

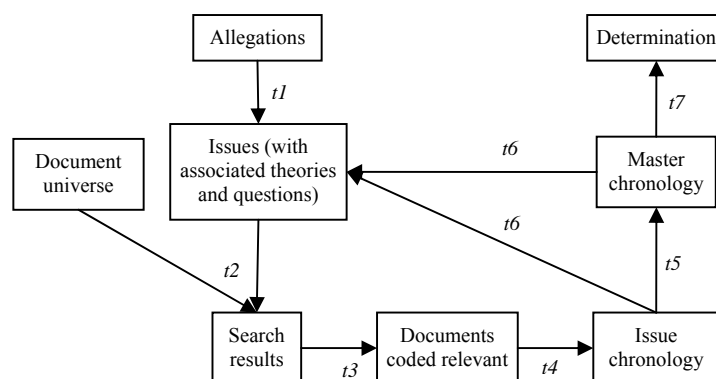


Figure 1. An overview of the investigation process.

representing an intermediate step that ultimately links allegation with a determination.

In overview, the transitions were:

t1 - Given the allegations, the investigators defined and recorded a set of issues that they wanted to investigate and associated questions they wanted to ask.

t2 – Given these questions, queries were submitted to the document universe to return documents relevant to each of the issues.

t3 - Returned documents were individually read and coded for relevance to the issues (within a document management system).

t4 – Relevant documents were then used to infer entries within issue-specific event chronologies.

t5 – Selected entries within separate issue chronologies were ‘escalated’ into a single master chronology designed to record the most significant aspects of the developing narrative.

t6 – By reflecting upon the narratives within the chronologies as they evolved, the investigators were able to identify apparent gaps, inconsistencies and periods of potential interest. This helped them to develop theories which guided the refinement of the investigation issues and associated questions.

t7 – given the knowledge acquired, the investigators formed a view concerning the allegations.

In fact, the structure of the investigation process evolved somewhat over time. What we present here is the process in its mature form. We have also restricted ourselves to a description of the investigation as it applied to electronic documents, omitting reference to witness interviews which were nevertheless an important, if non-technological, source of information.

A number of things are happening in this process, but broadly we see it as a process of information reduction achieved by different kinds of filtering and abstraction, directed by reflective interpretation on the part of the investigators.

Two things are important to note. First the investigators constructed each step for a reason—this being in general terms to help them move in a direction in which they

wanted to go. Hence we can learn about their needs from what they did. The second point is that although they had discretion to design the process as they saw fit, they did so within the constraints of the tools available to them at the time and whatever costs there were associated with their appropriation and use. Hence we can use the process to consider other tools which may have supported their needs better.

Focussing on transition t4

In considering where new technologies might offer leverage we might focus attention in detail on any part of the process we have described to consider how things might be changed (or even change the process as a whole). Our interviewees consistently cited manual document review (stages t3 and t4 in figure 1) as imposing the major overhead in terms of time and effort. Over the course of the investigation 130,000 documents were reviewed in all. This represents a significant reduction on the document universe, but is nevertheless a very significant number of documents. Here we will consider transition t4 in more detail.

T4 involved the creation of semi-standardized event records based on the review of documents which themselves had been coded as relevant to one or other issue. The investigators constructed chronologies as a table using Microsoft Excel according to a preformed schema. An example of an anonymised record which reflects this schema is shown in figure 2.

The reason for creating chronologies was so that the investigators could have a compact representation of events they considered to be of significance to their investigation. This then provided a resource for considering what they knew, for developing theories from this and in other ways establishing what it was they wanted to find out (transitions t6).

The resource for creating event records (transition t4) was a list of documents (predominantly emails) displayed in overview as a chronological file listing within a document management system. The task of the reviewer was to review each document in turn and, where appropriate, create a record of any event of potential significance to the investigation. For example, this might be a meeting proposed by email between one protagonist and another.

An appreciation of which events held significant to the

Date	Time	Event/Document	People Involved/ Author/Recipient	Evidence / File Reference
8th Nov	7.45	{company A} Meeting in {country A} (time is {person B} flight departure from {location A} to {location B}) with return to {location A} for 12.55 on 9th Nov. {person I} to pick up {person B} at Airport	{person I}, {person B} and {person H} in {location C}	Email between {person I}, {person I}, {person H} and {person F}/Doc ID 169246

Figure 2. An anonymised event entry from one of the chronologies

investigation (and hence what to record) evolved over time as the investigators' understanding developed and they reviewed their interests. What we focus on here, however, is what happens when an investigator first discovers information about a potentially significant event. The information contained in the message acts as a cue to the investigator about something that should be recorded. However, they are also aware that the information they have found is not the complete picture. For a meeting, the investigators we interviewed described a number of things they might like to know such as where and when it took place, who attended, what was discussed and what were the outcomes? Some or all of this information might be missing from the initial cue, and may be found distributed across a number of other messages. In addition, the initial lead may have been misleading: there may have been a change in plans or the meeting may not have actually taken place at all.

Following Klein et al's [2006] model of the process of sensemaking, we think of the investigator's concept of an event as an instance of a 'frame'. Frames are structures that we impose on the world in the process of understanding it. They are triggered by cues and act as plausible interpretations of those cues. A significant property of a frame which is important here is that they extend beyond the data from which they were cued. The ability to interpret things in this way is a fundamental human capacity. But as a consequence of this they can be wrong, perhaps as a result of a misleading cue.

Returning to the investigation, following the initial discovery or cue, a question arises about how to proceed. The initial document provides an important lead, prior to which the investigator knew nothing of the event. We may say that at this point they have a theory that a significant meeting took place. But this theory gives rise to a need for further information, specifically in order to address the need of elaborating and validating the interpretation.

In this situation the investigators we interviewed described two strategies. Given potential difficulties in locating other documents about a given event, one strategy was simply to record the event as a conjecture and move on. Investigators would raise an event record in a chronology (marked as a conjecture) and continue reviewing documents as before in the hope that they, or someone else, would come across further relevant information later. The second strategy was to construct further keyword and/or date delimited queries designed to re-filter the collection in a way that might bring relevant documents to the surface.

Whilst the second strategy offers continuity to the investigator in terms of focus by supporting a single chain of thought, it is also non-trivial. The investigator sees the situation under investigation in terms of events, whilst the system they are using sees the situation in terms of documents, textual tokens and metadata. Consequently, the investigator must translate their question (of all documents

relevant to a particular event) into something understood by the system—referred to more generally as a 'compromised need' [Taylor, 68]. This can require some cognitive effort and result in what is at best an approximation.

Reflections on design

This example suggests a general principle which we can apply to such problems. That is—where a user is making sense of information through the application of a particular type of frame (or frames), leverage can be obtained by linking information around possible instances of that frame (or frames) in the data. Of course, there may be a number of types of frame that are important to an investigator. Other frames we identified from our interviews included entire business activities (such as contracts), particular time periods surrounding major events within those activities, and protagonists or potential protagonists under investigation. In many of these cases, information discovered within the collection cued the investigators to their existence, raising them as foci for further investigation (the investigators started from an almost entirely blank slate). And in all cases further information was distributed across the document collection.

We note here that frames that were of significance to the investigators were reflected in the way the investigators structured the knowledge representations they generated i.e. chronologies structured in terms of individual events and individual chronologies dedicated to information about specific business activities and people. Hence, in understanding how the investigator seeks to translate information at transition t_4 , we may need to look no further than the representational form they seek to create.

Construing the investigation from the perspective of the investigator, then, it is a question of how frames are cued and how this leads to the need for their elaboration and/or validation. From this, we can ask the question of the extent to which information retrieval technologies support this thought process.

We have addressed this question to some extent in relation to the need for the investigator to translate their needs into terms understood by the system—terms which are characteristically low-level in their characterization of document content.

We may ask what other kinds of technologies might be developed which may be more helpful. The question here is one of raising the bar in terms of system intelligence in order to achieve the potential for aggregating documents in terms which more closely approximate to the concepts of the investigator. In this way, transformations performed earlier in the process (i.e. search) would organize the data in a way better adapted for subsequent work.

A number of possibilities exist here. First, systems that offer representations of email documents in terms of subject threads may offer some advantage. Analysis of the Enron collection, however, has suggested that the average length

of an email thread (in organizations at least) is typically quite short. Also systems that are capable of semantically clustering documents (e.g. Attenex) may be of value, depending on the extent to which emergent document clusters map to investigators' conceptual frames.

Another alternative is to use systems that perform information extraction (IE). IE system process free text and use techniques in computational linguistics in order to identify pre-defined elements of meaning [Gaisauskus & Wilks, 1998]. Jigsaw [Stasko et al], for example is an investigators tool specifically designed to graphically represent the results of information extraction over a free text collection. Elsewhere, capabilities for identifying temporal and event references in text have been demonstrated at 83% accuracy against hand-annotated data (Mani and Wilson, 2000).

DISCUSSION AND FUTURE WORK

We believe that a promising approach to the design of more appropriate systems for e-discovery work is to structure them around the terms or concepts in which the investigators understand the subject-matter of the investigation. The significance of these concepts is that they provide the vocabulary through which investigators see the world. The sensemaking process in e-discovery can be seen as one of translating large amounts of unstructured data into representations structured in these terms. The transitions represented in figure 1 can be seen as a process of filtering and abstracting information into these terms.

The approach we have illustrated involves the identification of the sensemaker's typical frames and the operations that they want to perform on them. By providing an analysis of related information needs and the way these develop, frames provide a foundation for reasoning about the design in terms of the typical cognitive paths users follow during sensemaking.

If systems can indeed be configured around the kinds of concepts that e-discovery investigators themselves apply to data, then they are likely to provide a higher platform on which investigators can apply their own expertise in sensemaking and allow them to work to a higher conceptual level [Rasmussen et al., 1994]. The ideal is that investigators can pursue investigations with fewer interruptions imposed by constraints of the systems that they use. And by identifying documents that are relevant to emerging concepts of the investigation, there is an opportunity to reduce the very high overhead of document review.

We intend to explore these ideas further in future work. We are about to embark on a further investigation case study. Of key interest will be the way in which investigators conceptualised their problem as expressed through the way that they talk about them and the ways in which concepts are rarified in the process of generating useful representations of knowledge.

We are also planning a laboratory study in which non-lawyer participants will perform a mock investigation using a subset of the Enron email collection. Manipulations in this study will involve the presentation of a document collection according to visual indexes based around different kinds of document aggregation, including email threads, semantic clusters and event references. Our aim will be to understand the value that these provide in the process of cueing, elaborating and validating users' conceptual frames.

ACKNOWLEDGMENTS

We would like to thank Freshfields Bruckhaus Deringer for their kind help with the fields study reported here. The work was funded under EPSRC grant EP/D056268.

REFERENCES

- Attfield, S., Blandford, A. & De Gabrielle, S (2008a) Investigations within investigations a recursive framework for scalable sensemaking support. Sensemaking Workshop, ACM SIGCHI Conference 2008.
- Attfield S., Fegan S. & Blandford A. (2008b) Idea generation and material consolidation: Tool use and intermediate artefacts in journalistic writing. *Cognition, Technology and Work* (online first).
- Baron, D. (2008) UK firms report jump in spend on e-discovery systems. *Computer Weekly*, March 2008
- Conrad, J.G. (2007) E-Discovery revisited: A broader perspective for IR researchers. DESI: Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings
- Gaizauskas, R. & Wilks, Y. (1998) Information extraction: Beyond document retrieval. *Journal of Documentation*, 54(1), 70-105.
- Klein, G., Phillips, J. K., Rall, E. L. and Peluso, D. A. (2006) A Data-frame theory of sensemaking. In *Expertise Out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making* (Pensacola Beach, Florida, May 15-17, 2003). Lawrence Erlbaum Associates Inc, US, 2007, 113-155.
- Mani, I. and Wilson, G. (2000) Robust processing of news. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000, Hong Kong)*, 69-76.
- Rasmussen, J. Pejtersen, A.M. & Goodstein, L.P. (1994) *Cognitive Systems Engineering*. New York, Wiley.
- Socha, G. & Gelbmann, T. (2007) The 2006 Socha-Gelbmann electronic discovery survey report. Socha Consulting, Saint Paul, MN.
- Stasko, J., Gorg, C. & Liu, Z (2008) Sensemaking across text documents: Jigsaw. Sensemaking Workshop, ACM SIGCHI Conference 2008.

7

Jigsaw: Investigative Analysis on Text Document Collections through Visualization

Carsten Görg and John Stasko
School of Interactive Computing & Gvu Center
Georgia Institute of Technology
Atlanta, GA 30332
{goerg,stasko}@cc.gatech.edu

ABSTRACT

This article describes the Jigsaw system for helping investigative analysis across collections of text documents. Jigsaw provides multiple visualizations of the documents and the entities within them to help investigators discern embedded stories and plots. Our early focus within Jigsaw has not been on legal documents and E-discovery, but we feel that the system may have potential in these areas as well. This article illustrates Jigsaw's views and operations using Enron email archives as example documents.

Author Keywords

Sensemaking, investigative analysis, information foraging, information visualization, multiple views.

INTRODUCTION

We have been developing a system called Jigsaw to help investigative analysts explore and make sense of collections of text documents. In particular, we have designed Jigsaw to help investigators uncover stories, plots, and threats embedded across the documents. While our focus has not been on legal documents or E-discovery, we are curious to explore whether Jigsaw might be useful in these areas as well. This article provides a brief overview of Jigsaw and its capabilities, using a subset of the Enron email archive as an example document collection.

Jigsaw has been developed to help people with sensemaking, exploration, and analysis activities on collections of unstructured, plain text documents, in particular, relatively short documents (approximately 1-10 paragraphs) in loose narrative form. Examples of such documents include police case reports, short news articles, or email notes. While Jigsaw can process longer documents, its utility degrades in these cases (reasons will be illustrated later in the article). The email shown below, taken from the Enron data set, is a good example of the kind of document ideal for Jigsaw.

Email 114844

Message-ID: <22094025.1075842958662.
JavaMail.evans@thyme>
Date: Fri, 1 Sep 2000 00:43:00 -0700 (PDT)
From: steven.kean@enron.com
To: jeff.dasovich@enron.com,
susan.mara@enron.com, mona.petrochko@enron.com,
tim.belden@enron.com, mary.hain@enron.com
Subject:

Cc: paul.kaufman@enron.com,
richard.shapiro@enron.com,
james.steffes@enron.com
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: paul.kaufman@enron.com,
richard.shapiro@enron.com,
james.steffes@enron.com
X-From: Steven J Kean
X-To: Jeff Dasovich, Susan J Mara,
Mona L Petrochko, Tim Belden, Mary Hain
X-cc: Paul Kaufman, Richard Shapiro,
James D Steffes
X-bcc:
X-Folder: \Jeff.Dasovich-Dec2000\Notes Folders
\All documents
X-Origin: DASOVICH-J
X-FileName: jdasovic.nsf

When we have described the problems and solutions for California we have focussed on generation siting and flexibility to hedge. We have stayed away from transmission issues on the assumption that California, with its ISO and PX, does not suffer from the same discrimination issues as other parts of the country. Is this true? Does California's system layer in priorities for utility use of the system -- eg doesn't PG&E control "path 15"? Does that control provide advantageous access to PG&E? Are there other examples and are there links between these "preferences" and the current problems in California? As we are trying to convert reliability and pricing concerns into FERC action these would be helpful arguments to have available to us.

Investigators may seek to connect the individual facts and events described in specific documents into a larger, more coherent thread or story. Putting the pieces together in this way can lead to a better understanding of the broader, more general notions and implications of the document collection.

Jigsaw's particular focus is on illuminating connections between the entities in the documents—the people, organizations, places, and so on. Jigsaw visualizes the documents

and the entities within them in a number of different representations, each one specifically created to communicate some different aspect of the data. For instance, Jigsaw can help to understand social networks of people, connections between people and places, and the evolution of events in time.

Within our research communities, these types of activities are known as sensemaking [3, 4, 5]. Investigative reporters, law enforcement officials, and intelligence analysts all routinely perform these types of activities. Clearly, as the number of documents being examined grows, the sensemaking activities become more challenging.

A variety of approaches to support people in sensemaking scenarios like the ones we describe do exist. Some use automated techniques and tools that examine a document collection without human intervention and report on discovered plots or narratives. These approaches typically use techniques and algorithms from the fields of artificial intelligence, data mining, and machine learning.

Our approach is quite different, instead involving human-centered investigations where we provide human analysts with computational tools to assist them while conducting investigations. Our tools seek to enable the powerful perceptual capabilities of people and bring those capabilities to bear throughout the sensemaking process. We firmly believe that human analysts harbor tremendous investigative skills, but the masses of data and documents typically present today can overwhelm the analysts' investigative capabilities. Thus, we provide visualization tools that transform the data (text documents in our case) into visual representations that can more easily be surveyed, scanned, examined, reviewed, and studied.

In order to facilitate the powerful exploratory, investigative skills of people, our tools are highly interactive and flexible. We seek to help analysts browse the document collection rapidly and to more deeply explore "interesting" avenues of investigations. Analysts must uncover whether the agents and events in question relate to potential plots being developed. Our approach also hinges upon multiple visual representations of the documents and entities within them. Any one visualization simply may not provide the right perspective onto the data to allow an analyst to perceive an important connection. By supplying multiple visual representations of the data, each providing a view onto some important characteristic, we are more likely to help the analyst discover the unknown connections that weave a larger narrative together.

Thus, Jigsaw espouses an *information visualization* [1, 6] approach to investigative and exploratory data analysis. More specifically, when visual approaches like this are combined with computational techniques to manage and filter the extremely large data sets that may be present, the resulting system illustrates *visual analytics* [8] principles.

Clearly, many different types of investigations occur within the legal community. Realistically, we speculate that Jig-

saw's utility is limited for typical E-discovery type tasks that may involve millions of documents. Instead, Jigsaw's value likely rises when a document collection has been narrowed down to a few thousand documents and an investigator wants to understand how the people, organizations, and events in those documents interact to reveal the "big picture."

JIGSAW

Jigsaw [7] is a system for helping analysts with the kinds of investigative scenarios discussed above. It is a multi-view system, including a number of different visualizations of the documents in the collection and the entities (people, places, dates, organizations, etc.) within those documents. Accompanying the visualizations is a textual search query interface so that particular entities can be examined directly. When used in this way, Jigsaw acts like a search engine that simply displays results through visualizations rather than text lists.

Jigsaw is much more than a search engine with visual results, however. Once views show documents and their entities, users can explore the collection by interactions with those objects. For instance, new entities can be displayed and explored by simple user interface operations in the views that expand the context of entities and documents. In fact, far more entities and documents are initially displayed via user interaction than by textual search queries. Search queries often serve to jump-start an exploration, but view interaction yields richer representations and exploration.

Most of the views in Jigsaw illustrate connections between documents and entities or between entities and other entities. Jigsaw uses a simple model of "connection" — an entity is connected to a document if it appears in that document (and vice versa) and two entities are connected if they appear in at least one document together. Entities that appear in more than one document together are considered to be more strongly connected with the connection value dependent on the total number of documents of co-occurrence. This simple model of connection is easy to implement, is easy for people to understand, and we have found it to be powerful for helping exploration of document collection.

The views in Jigsaw are linked so that actions in one view propagate to the other views whose visual state updates to reflect that action. For example, the most common operation on a view is to mouse-click on an entity or document which selects that object, and then the rendering of other objects in the view updates to reflect their relation to the selected object. In Jigsaw this action is propagated to other views which then also select that same object and update their displays appropriately. Another common operation is to "expand" an entity or document which typically displays a new set of entities and documents that are connected to this object. This operation is usually invoked by a double-click on an object or a click-activated menu.

The person using Jigsaw also can decouple a view from event listening so that its visual state only changes via explicit operations in that view. We have found this capability to be very useful when an analysis process yields a view configu-

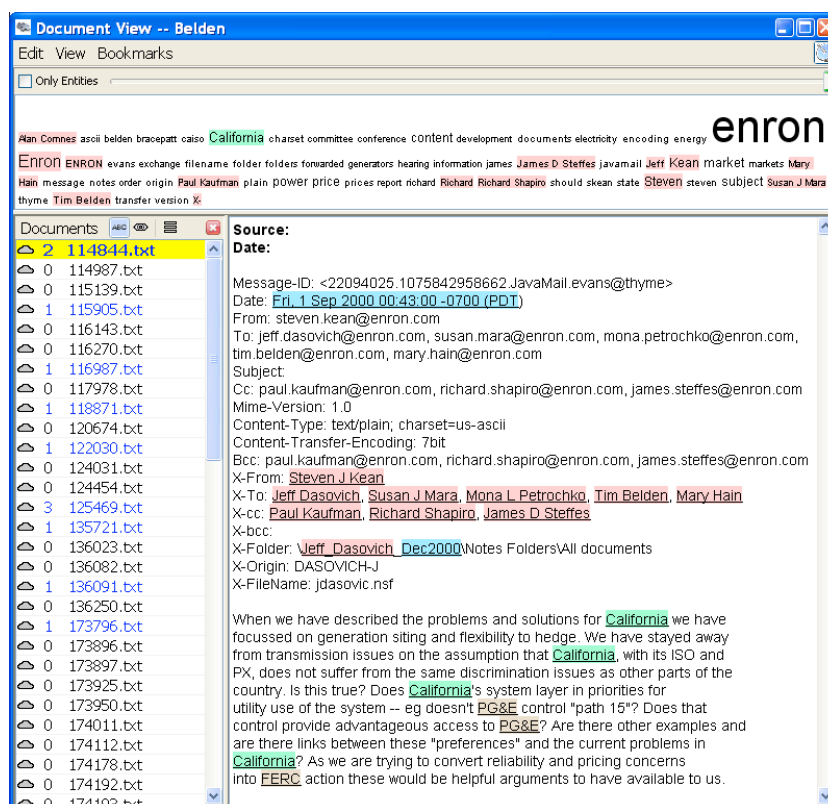


Figure 1. The Document View showing an example report.

ration that is particularly enlightening and the user wants to keep the view as-is during subsequent exploration.

Jigsaw's views include list, graph and scatterplot-based representations of object connections, an overview-style document cluster view showing all documents, a calendar view for examining temporal patterns, and a fundamental document view showing document text with highlighted entities. Below we describe some of these views in more detail.

In Figure 1 the Document View shows the example document mentioned in the introduction. To facilitate fast scanning of text documents, entities are highlighted according to their type. The tag cloud at the top of the view describes the contents of the marked documents in the document list.

Figure 2 shows the connections of "Tim Belden" in the List View. For each of the lists an entity type can be selected and the lists can be sorted either by frequency, alphabetically, or by the strength of the connection. The bars on the left border of each list entry display the frequency across the whole document collection of the entry. Connections between entities are visualized in two different ways: items connected to a selected entity are marked in a shade of orange (the stronger the connection, the darker the shade of orange) and in neighboring lists connected entities are additionally joined by lines. Thus, it is possible to see which entities are connected in case multiple items are selected.

Figure 3 shows the Graph View. The larger white rectangles represent documents, the smaller colored circles represent entities (colored according to their type). By expanding and collapsing nodes to either show or hide their connected entities or documents respectively, the analyst can explore the network step by step.

Figure 4 shows the Calendar View. Documents and entities from the data set are displayed in the context of a familiar calendar showing years, months, weeks and days. The small diamond items drawn on a particular day represent documents (colored gray) or entities (colored according to its type) in the context of the date(s) noted in documents in which they appear. When the user moves the mouse pointer over a document-representation diamond drawn in the calendar, all the entities appearing in that document are shown on the left.

We have found that the system is more useful when a set of views can be laid out and easily examined without window flipping and reordering. Due to the large amount of screen real estate required to display its views, Jigsaw ideally should be run on a computer with multiple and/or high-resolution monitors.

More details about Jigsaw can be found in [7] and at the project website:

<http://www.cc.gatech.edu/gvu/ii/jigsaw>.

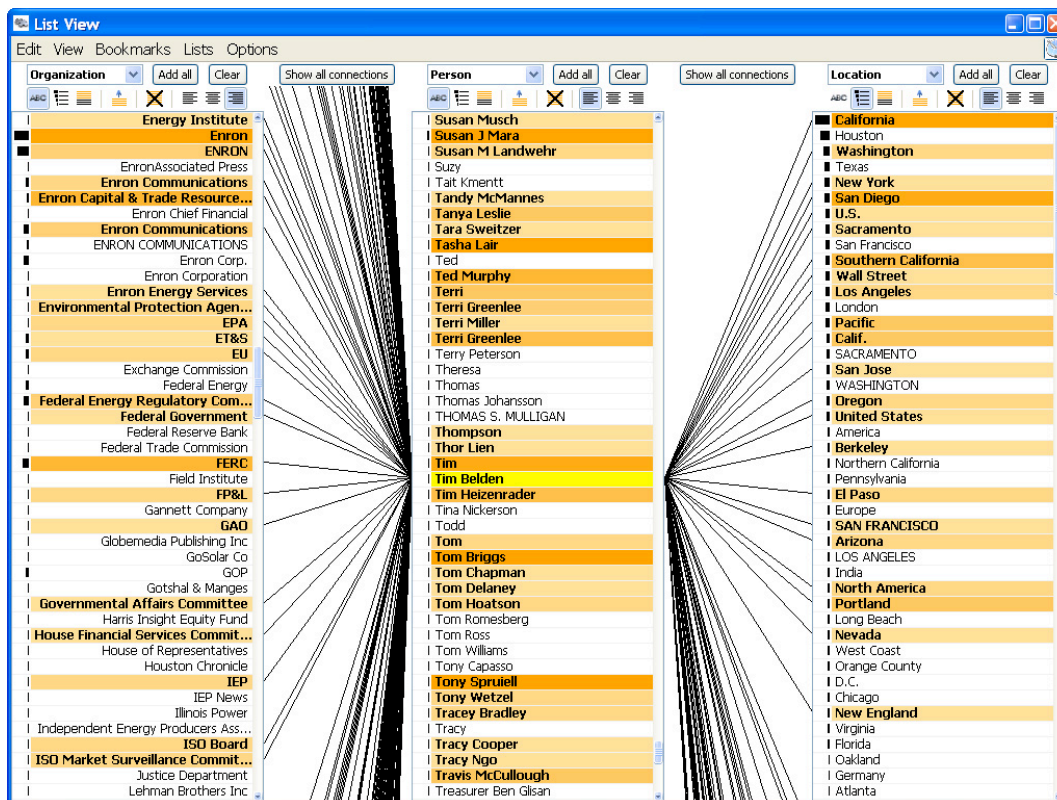


Figure 2. The List View showing connections of “Tim Belden”.

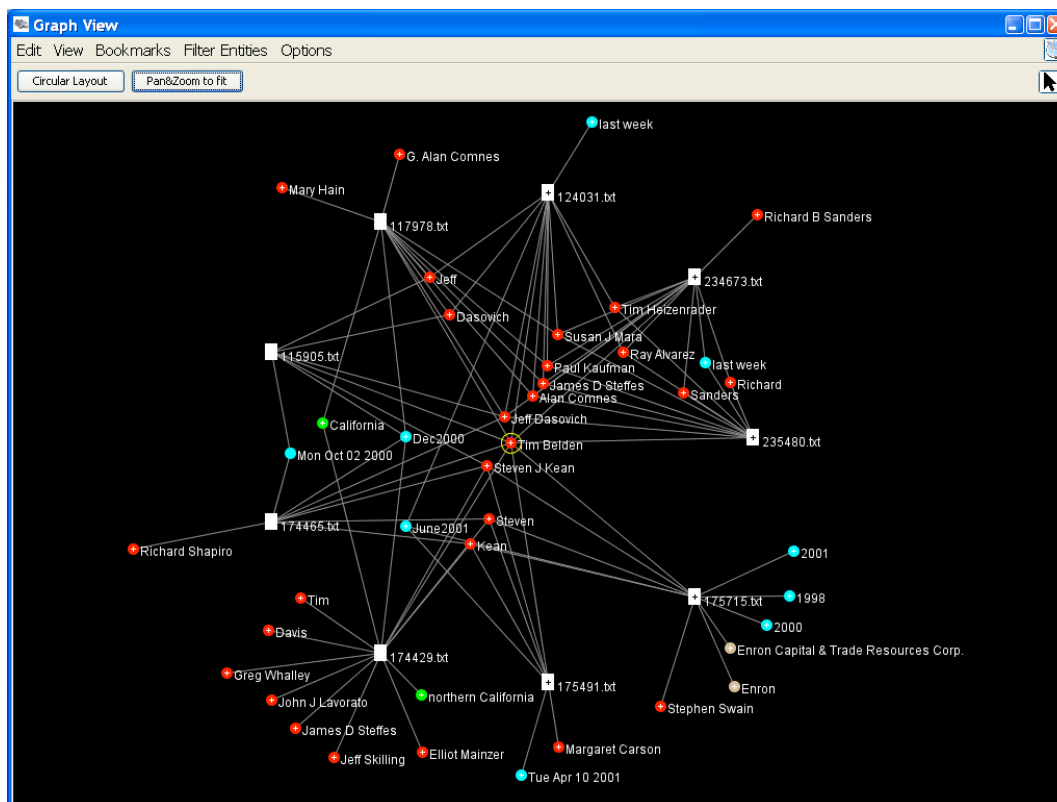


Figure 3. The Graph View after exploring some connections of “Tim Belden”.

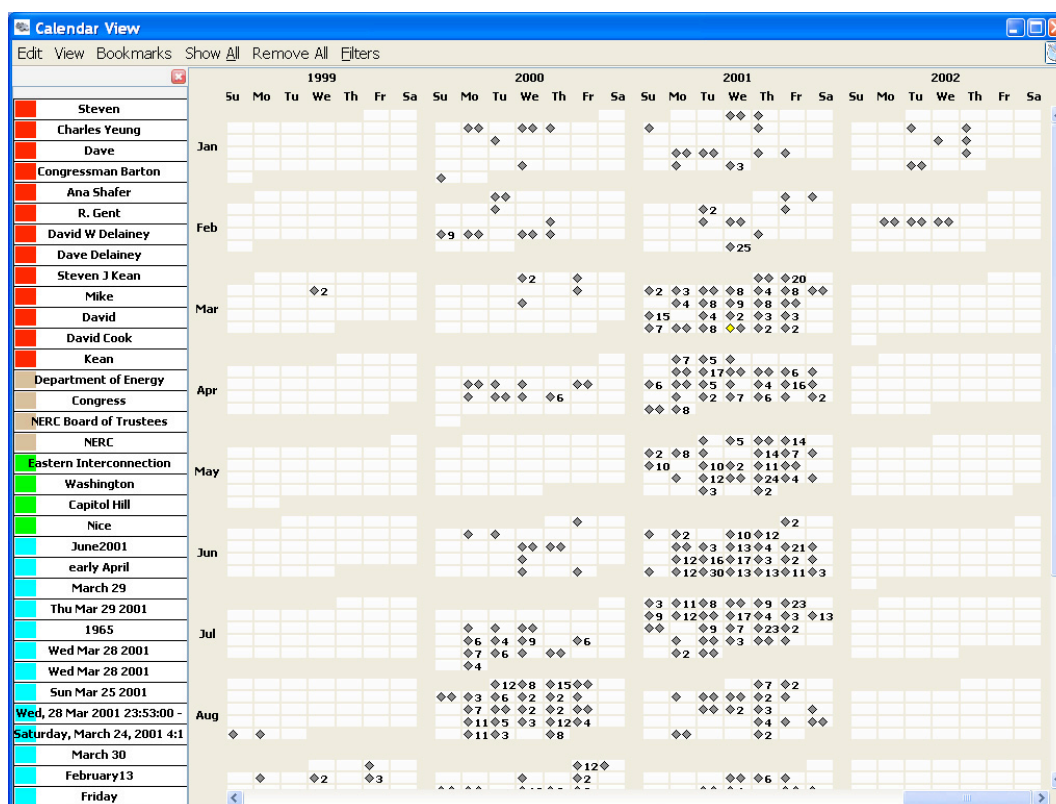


Figure 4. The Calendar View showing when emails were sent.

We want to stress that Jigsaw does not seek to depict the main themes running throughout the document collection or the semantically meaningful concepts within it (although these are worthy goals for future work). Presently, Jigsaw acts like a visual index onto the document collection, helping to provide fast, contextualized access to the individual entities and documents that an analyst is studying.

Fundamentally, an analyst must read documents to understand the events occurring within them. As document collections grow larger and larger, finding the most fruitful documents to read becomes more challenging. Furthermore, traditional search technology is not as useful in this situation because the plots/stories discovered often involve unexpected and serendipitous connections between entities which are best found following a trail of linked evidence.

SENSEMAKING ACTIVITIES

In terms of the sensemaking model proposed by Pirolli and Card [4], we feel that Jigsaw can help analysts with both the information foraging and sensemaking loops, but its utility is much stronger for foraging right now. As discussed above, Jigsaw helps people find small collections of potentially important documents to read and study, a fundamental activity in information foraging.

To support the evidence marshalling and sensemaking process, Jigsaw provides a special view called the Shoebox. The Shoebox helps the analyst to collect and organize items or

information of interest that were revealed while exploring the document collection. Figure 5 shows an example of the Shoebox view.

The analyst can add items to the Shoebox from every view — they appear first in the ‘inbox-area’ on the left side of the Shoebox. Items added at the same time are grouped together and sorted by type. The Shoebox offers multiple ways to organize the items in the inbox and to join them to build sensemaking artifacts:

- Combining items to sentences by adding comments and snapping entities together
- Grouping items according to a topic
- Forming hypotheses and using items as supporting or contradicting evidence
- Linking hypotheses, groups, sentences, and items.

These sensemaking artifacts support the analyst’s thinking process in a visual way and reduce the amount of necessary text as much as possible. This is important since the analyst may already be overwhelmed with text documents. During an informal evaluation of Jigsaw with an analyst, a reoccurring statement was: “I don’t want to read it, I want to see it.”

While designing the marshalling support for Jigsaw, we envisioned two different approaches for collecting evidence:

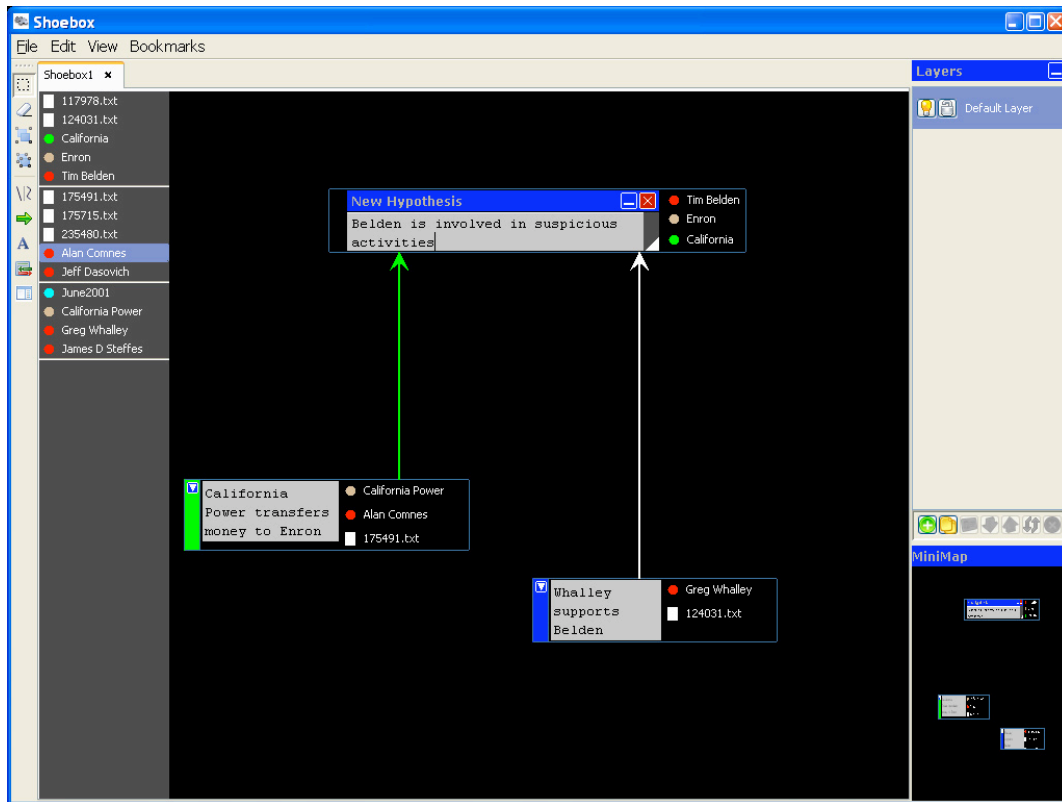


Figure 5. The Shoebox View showing the hypothesis, group, and link feature.

either augmenting the existing (data) views, or collecting evidence in a separate Shoebox view. Incorporating the marshalling process into the existing views would have the advantage of placing the information along with necessary comments right at the spot where it was discovered rather than duplicating information at another location. The disadvantage would be that evidence would be scattered across multiple views what would make it difficult to keep track of the collected information. Therefore, we decided to collect evidence in a separate Shoebox view. To address the problem of duplicating information, we added a hyperlink function to the Shoebox. This allows the analyst to connect views via links to bookmarks as evidence to the Shoebox.

CONCLUSION

In this article we have described the Jigsaw system that has been designed to help investigative analysts find embedded plots or stories in large document collections. We believe that this type of exploration may be useful in legal activities where sensemaking and analysis occur.

Jigsaw provides multiple visualizations of documents and the entities within them, as well as the connections that exist between entities and/or documents. Jigsaw provides a decidedly human-centered approach to sensemaking by allowing people to interact with the views and explore possible new avenues of examination. Presently, the system provides more information foraging utility than schema/hypothesis generation utility, but we are exploring how these latter ca-

pabilities could be added to the system too.

Evaluation of Jigsaw is an ongoing activity as well. Presently, we are conducting experiments to examine whether people can use individual views to answer the kinds of analytic queries common to the domains we study (e.g., Do these two people share any common acquaintances? Has this person ever been to that city?) Our next evaluation phase will involve more holistic study of the system to see if it does benefit analysis as compared with investigations using more common aids such as search engines and authoring/organizational tools. To do that, an analysis activity may have to be conducted over days rather than minutes. Finally, the utility of Jigsaw was illustrated at least informally by our use of the system to win the university component of the 2007 IEEE VAST Symposium Contest [2].

ACKNOWLEDGMENTS

This research is supported in part by the National Science Foundation via Award IIS-0414667 and the National Visualization and Analytics Center (NVACTM), a U.S. Department of Homeland Security Program, under the auspices of the Southeast Regional Visualization and Analytics Center.

REFERENCES

1. CARD, S. K., MACKINLAY, J., AND SHNEIDERMAN, B. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, 1999.

2. GÖRG, C., LIU, Z., PAREKH, N., SINGHAL, K., AND STASKO, J. Jigsaw meets Blue Iguanodon - The VAST 2007 Contest. In *IEEE Symposium on Visual Analytics Science and Technology* (2007), pp. 201–202.
3. KLEIN, G., MOON, B., AND HOFFMAN, R. Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems* 21, 4 (July 2006), 70–73.
4. PIROLI, P., AND CARD, S. Sensemaking processes of intelligence analysts and possible leverage points as identified through cognitive task analysis. In *2005 International Conference on Intelligence Analysis* (May 2005).
5. RUSSELL, D. M., STEFIK, M. J., PIROLI, P., AND CARD, S. K. The cost structure of sensemaking. In *CHI '93: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (New York, NY, USA, 1993), ACM, pp. 269–276.
6. SPENCE, R. *Information Visualization*. ACM Press, 2001.
7. STASKO, J., GÖRG, C., LIU, Z., AND SINGHAL, K. Supporting investigative analysis through interactive visualization. In *IEEE Symposium on Visual Analytics Science and Technology* (2007), pp. 131–138.
8. THOMAS, J. J., AND COOK, K. A. *Illuminating the Path*. IEEE Computer Society, 2005.

8

Reconstructing Financial Statements

Frank G. Bennett, Jr.¹

Nagoya University, Nagoya 464-8601, Japan,
`bennett@law.nagoya-u.ac.jp`,

WWW home page: <http://gsl-nagoya-u.net/faculty/member/gslF.Bennett.html>

Abstract. This paper introduces a tool for the reconstruction and validation of categorized totals embedded in untrusted and unformatted text, such as OCR scans of financial statements. The tool is a spinoff of academic research into the funding of Japanese third-sector organizations, the annual reports of which are frequently published reports in the form of PDF files containing document images. A number of techniques at string- line- and document-level are used to resolve ambiguities and obtain the greatest possible recovery rate for the underlying data, while excluding the content of untrustworthy documents from the final sample. In a preliminary trial “in the wild”, the tool has returned validated income totals for 47.9% of the documents in a heterogeneous set of 2205 annual reports.

1 The Problem

In common with litigation, empirical research in Japan is at times afflicted by what one might call the “last byte nuisance”, whereby information is disclosed in forms that are costly to analyze in bulk. In the last century, it was the common practice, when disclosure was required by administrative law, for government agencies to offer documents for reading at a single location, without provision (sometimes without permission) for copying.¹ Such practices have their modern corollary in the copy-protected, print-restricted PDF wrapper for documents presented in graphical form.

While PDF content restrictions are trivial to handle, and OCR tools can produce an electronic text version of a document, raw OCR output cannot be trusted as a basis for analysis. This is a particularly telling issue in the case of financial data, where small recognition errors can greatly affect the apparent meaning of document content. In this case, means of both reconstructing the essential structure of the document, and of validating the figures themselves can greatly increase the value of such data as a basis for answering research questions. This paper presents a tool designed for this purpose. While the incentives for the drafting of this tool are “entirely academic”, the methods outlined here

¹ A particularly odd example of such “managed transparency” is the now defunct practice of restricting notetaking by spectators to court proceedings. *See, e.g.* Supreme Court judgment of Mar. 8, 1989 (Case of note-taking in court without permission), 1299 HAREI JIHO 341.

can be applied to any large archive of similar data, and the source code of our implementation is available for download under an open source license.²

The paper begins with an introduction of the research context, to provide a sense of the development environment and the objectives of the researcher. This is followed by an outline of the data collection infrastructure into which the tool fits, and brief technical overviews of programming logic at the string, line, and document levels of processing. The paper concludes with comments on the results of initial testing, and the on the focus of future development.

2 The Research Context

Japanese rules on non-profit corporations changed significantly under law reforms taking effect in 1999 and in 2008. Prior to this time, most non-profits depended on specific approval by a national-level ministry as a condition of their creation. The first significant loosening of this restriction came with legislation defining a new legal person (the so-called “NPO law”), suitable for small or volunteer-driven associations, which can be created with no investment of capital, upon the satisfaction of a set of objective requirements. The legislation has been well received, and this year the advance approval requirement was removed from all non-profit entities.

The Japanese third sector is currently very small, but expanding at a much faster rate than the surrounding economy. The first 10 years of the NPO law have seen the founding of 35,000 entities, with an average of about 15 more being added every working day. Following the recent extension of favourable tax treatment to qualifying non-profits, we can expect this trend to gather greater force in future years.

Non-profit organizations can be used for various purposes. In many industrialized societies, they provide a platform for the lobbying of politicians and bureaucrats. In Japan they have heretofor served, on the contrary, as an extension of government administration. Such organizations can also provide an internal support network to their membership, with a degree of independence from government and the surrounding community. The mere fact that the number of these organizations is on the rise therefore tells us little about the impact that they have on government and society. [1] To explore this research question, we need to explore what makes them tick, and the obvious way to do so is through their financial reports.

Each non-profit is required by law to file a financial report each year with the government authority (local or national) responsible for its incorporation. Government is required to archive these reports for a period of three years, and to make them available to members of the public on request. Most of the relevant government bodies fulfill this requirement in the traditional way described above, by providing a reading room where concerned citizens can request and view the documents in paper form.³ But a significant number of local authorities now

² The URL for download is <http://gsl-nagoya-u.net/appendix/software/renumerate>.

³ See, for example, the disclosure policy of Metropolitan Tokyo. [2]

provide these annual reports via the World Wide Web, in the form of graphical images encased in a PDF wrapper. These latter documents are the primary target of the digitization strategy discussed here.

Fig. 1. Sample double-entry financial statement

特定非営利活動法人 ウエルネスサポート

収支計算書

平成18年 4月 1日～平成19年 3月31日

平成18年度 特定非営利活動

(円)

科目	予算額	決算額	差額
I 収入の部			
1 基本財産運用収入	0	0	0
2 入会金・会費収入	0	0	0
3 事業収入	230,000,000	231,882,343	1,882,343
介護保険収入	230,000,000	231,882,343	1,882,343
4 補助金等収入	1,000,000	978,882	-21,118
雇用助成金	1,000,000	978,882	-21,118
5 雑収入	310,000	304,060	-5,940
受取利息・配当金	10,000	5,364	-4,636
雑収入	300,000	298,696	-1,304
6 借入金収入	0	0	0
7 その他収入	1,700,000	1,775,374	75,374
預り金等増加収入	1,700,000	1,775,374	75,374
当期収入合計(A)	233,010,000	234,940,659	1,930,659
前期繰越収支差額	15,929,848	15,929,848	0
収入合計(B)	248,939,848	250,870,507	1,930,659
II 支出の部			

3 The Documents

Each document in the target set is in one of two common formats, referred to here as the *double-entry* and the *running-totals* formats. Samples given in Figures 1&2 show the respective structures of these document types. In both types of income statement, totals are embedded in the series of figures, reading the document from left to right and top to bottom. In *double-entry* format, a total occurs at the end of each line. In *running-totals* format, totals occur at irregular intervals down the page, with a grand total at the bottom of the run of numbers.

These patterns are obvious to the human eye, but as Figure 3 illustrates, much of the information on which a human reader relies can be lost in OCR output text. Our aim is to exploit embedded totals as a checksum hint for the repair of OCR damage and the reconstruction of the original document content.

Fig. 2. Sample running-totals financial statement

2006年度特定非営利活動に係る事業会計収支計算書			
2006年4月1日から2007年3月31日まで			
特定非営利活動法人 〇〇〇〇〇〇〇〇〇〇			
科 目	金 額 (単位 円)		
(資金収支の部)			
I 経常収入の部			
1 会費収入			
正会員(個人・法人)	1,360,000		
賛助会員(個人・法人)	0		
認定講師	495,000	1,855,000	
2 事業収入			
技能検定事業収入	11,182,000		
福祉活動事業収入	0		
調査・研究事業収入	1,185,000		
研修・講座・講演事業収入	2,107,468		
広報事業収入	288,660		
その他事業収入	0	14,763,128	
3 基本金運用収入			
基本金利息収入	964	964	
経常収入合計			16,619,092
II 経常支出の部			
1 事業費			
技能検定事業費	6,277,297		
福祉活動事業費	0		

4 The Recognition Engine

In our specific project, we have settled on the *Tesseract* OCR engine for use in the first step of the processing chain. This product was developed by Hewlett-Packard between 1985 and 1994, [4] was later released under an open source license in 2005, and is now under active development at Google. [5], [9] Alone among open source systems, *Tesseract* has been subjected to rigorous competitive testing, ranking third overall in a test of eight leading contemporary systems carried out by the Information Science Research Institute in 1995. [3]

Tesseract is fully trainable and is able to achieve a high recognition rate against heterogenous text. The output against a sample document after training is shown in Figure 3. Two features of this target data will be immediately apparent. Most obviously, much of the page layout information has been lost. Furthermore (as Japanese readers will immediately note), the system recognizes only a limited set of Japanese characters, plus digits and symbols found in the numeric data. In fact, there is an upside to both of these limitations.

Page layout information is extremely useful when interpreting documents with a known homogenous structure, but the target documents in this case are in varying formats. The position of a number on the page, cannot be used as a primary hint to the significance of a particular number. Loss of layout information forces us to concentrate on the pattern of sums in the number series, which is a more certain means of validation. The limited scope of recognition is also

a beneficial constraint, in that it avoids some of the ambiguous glyphs, such as “one” and “lowercase letter *el*” (1 and l), or “zero” and “capital letter *oh*” (0 and O), that plague OCR systems operating on English/Roman documents.

The Japanese phrases used in our early recognition trials are presented in Table 1. In all, the engine was trained for a total of 157 common characters and symbols, including the digits zero through nine. Because *Tesseract* is aggressive about recognizing individual character blobs even at very low confidence levels, providing additional “noise” characters in the training set reduces the possibility of false positives when post-processing the OCR output, while staying well within the capacity limits of the current version of *Tesseract*. The boxed areas highlight successful recognition of target phrases, and are used to identify the category of individual items of income during processing.

Table 1. Minimal character set for OCR training

Characters	Romanization	Translation
収支	<i>shūshi</i>	income and expenses
会費	<i>kaihi</i>	dues
事業	<i>jigyō</i>	project
寄付	<i>kifu</i>	gift
助成	<i>josei</i>	grant (private)
補助	<i>hojo</i>	grant (public)
利息	<i>risoku</i>	interest
經常	<i>keijō</i>	ordinary
収入	<i>shūnyū</i>	income
合計	<i>gōkei</i>	total
支出	<i>shishutsu</i>	expenses

5 Post-Processing

To prepare the text for string-level analysis, a series of regular expression substitutions are applied to normalize the text, repairing “impossible” character combinations. In preparation for later line-level analysis, the document is split into individual lines, and lines are fed to line-level resolvers. Line resolvers split the line into the string-level units that form the basis of document resolution.

Processing at the string, line and documents levels is outlined below. While the author has no formal training in computer science, the discussion should be accessible to readers knowledgeable in C++, Java, and other object oriented languages.

Fig. 3. Result of recognition (boxes and explicit space markers added)

1	2006年度特定非営利活動に係る 事業 会計 収支 計算書
2	2006年4月1日から2007年3月31日まで
3	特定非営利活動5ま人日事】 (一・) 十ル方5一構会
4	科目額(単位円)
5	(>収の,)
6	1経常 収入 の部
7	1 会費 収入
8	年会費(1団人・5ま人) 1,360,000
9	費助会費(借人・賃五人) 0
10	書講定講日市 495,000, 1,855,000
11	2 事業 収入
12	講講構定 事業 収入 11,182,000
13	補業位活動 事業 収入 0
14	講費・町開 事業 収入 1,185,000
15	町借・講座・講講 事業 収入 2,107,468
16	座講 事業 収入 288,660
17	受の地 収入 0,14,763,128
18	3講ま金道開 収入
19	講取金 利息 収入 964,964
20	経常 収入 合計 16,619,092
21	0経常 支出 の部
22	1 事業 費
23	位講構定 事業 費 6,277,297
24	補業ま活動 事業 費 0

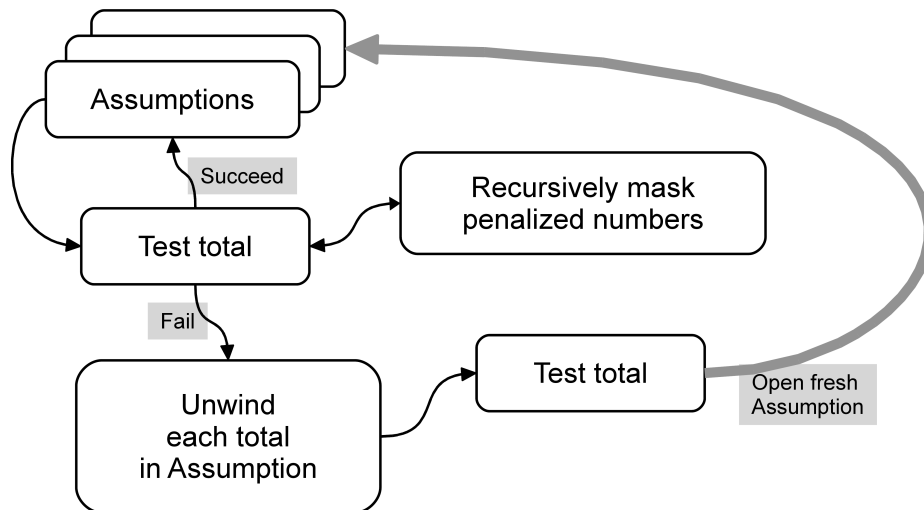
5.1 String Level Resolution

After normalization, each line is scanned for relevant strings with a regular expression parser to extract a list of relevant string elements. These elements (for example, at line 12 of Figure 3, the strings **事業**, 11, and 182,000) are then identified as either relevant text strings (such as **事業**) or numeric strings (11 and 182,000). Relevant text strings are associated with a category, and this is recorded for reference.⁴ The text string itself is discarded.

Numeric strings come in two flavours, depending on whether they can be interpreted as a single unambiguous unit, or have multiple elements that introduce ambiguity. For simplicity of processing, all numeric strings are first cast as

⁴ A record of the current category is maintained in a `categoryHinter` state object that persists across lines in the splitting process.

Fig. 4. Reconstructing totals within data stream



`ambiguousCluster` objects, and each contiguous run of numbers and commas within them is instantiated as an `ambiguousNumber`. In the example at line 12 of Figure 3, the numeric strings 11 and 182,000 constitute an `ambiguousCluster` composed of two `ambiguousNumber` objects.

The `ambiguousNumber` class is the fundamental unit for subsequent processing, and carries important metadata used in the remainder of the resolution process, and in final reports generated from validated number sets.

5.2 Line Level Resolution

After creating `ambiguousNumber` objects and wrapping them in `ambiguousClusters`, an attempt is made at the line level to resolve ambiguities in multi-element clusters. The version of the tool used in our project applies two line-level resolvers in separate attempts at “disambiguation”. The first applied is the *double-entry* resolver. As implemented for the first trial, this operates exclusively on `ambiguousCluster` objects containing three or more numbers, based on the hypothesis that three numbers exist in the line, and that the third represents the difference of the first two. This resolver always returns either the second number in this sequence, or an empty line. (There was a difficulty with this assumption, as explained in the conclusion.)

The second resolver, aimed at the *running-totals* format, operates on all numbers in the line, on the hypothesis that any multi-element `ambiguousCluster` could consist of either one or two numbers. In the latter case, the second number would represent the total of the preceding element, plus some unknown number of elements on previous lines of the document. Multi-element clusters are therefore

joined, beginning from the last item, such that the value of the joined number is *not greater than* the value of its predecessors joined. This resolver always returns all values in the line; but because the algorithm exits when the not-greater-than condition is reached, the resulting line may still contain unresolved multi-element **ambiguousCluster** objects when it is passed on for document-level resolution.

5.3 Document Level Resolution

Document level processing is the final stage of resolution. At this stage, all potential ambiguities in the text must be excluded, with the return of a single result representing the numbers in the original document and their characteristics. The tool returns this data in the form of a list of **ambiguousNumber** objects, in the same order as in the original document, with the income category and the number's role as an ordinary numeric component, total or grand total stored on the object as metadata. From this final list, spreadsheets and other reports can be generated.

In the case of *running-totals* line resolution, ambiguities may persist in the set of numbers returned for document level resolution. As more refined methods have been exhausted at this point, the tool now resorts to the delicate application of brute force. One complete, unambiguous series of numbers is produced for every possible combination of joins within the multi-element clusters remaining in the document. These number sets are the basis for final resolution of the document.

The preferred method of resolution at the document level is the mathematical pursuit of totals. The tool steps through the numbers in each **Assumption**, maintaining a running total of their values. When a number is encountered which equals this total, that number is treated as a **total**, and its value is added to a running grand total. When a final number is encountered which matches this grand total, its component **ambiguousNumber** objects are composed into a list of **candidates**, and processing of that **Assumption** finishes.

At significant values, the possibility of a false grand total is so small that it can be ignored for our purposes. However, at very small values, false totals are sometimes returned. After the processing of all **Assumptions** is complete, the tool returns the set of numbers for the **candidate** which results in the largest grand total. Because this recovery method can be foiled by OCR corruption of the numbers constituting the true total, very small values (those less than 5000) are excluded from the result.

Two further potential sources of failure or error remain. As shown in Figure 1, an income statement often contains numbers irrelevant to the finance figures to be extracted. Not all of these can be safely excluded at the initial stage, before **ambiguousNumber** objects are instantiated. Their presence will block identification of the correct total. To address this problem, each **ambiguousNumber** object is affixed with a **penalty** value based on its value and its position within the document.⁵ After each failed attempt to achieve a valid total, the number with the lowest negative penalty value is discarded, and the totals process is retried.

⁵ The assignment of penalties is controlled by the **penaltyEngine** class, which can be customized to fit the characteristics of a given document set.

A second issue is the possibility of falsely identified totals within the number set. For example, in the sequence 111, 222, 333, 666, 666, the grand total of 666 will not be identified if the third number in the sequence is treated as a subtotal (in that case, the last 666 would be recognized only as a subtotal of the single number before it). To obtain the best internally consistent set of totals, each time a total is identified, an additional **Assumption** is generated, in which the same number can only be treated as a non-total value. This alternative **Assumption** is placed in the processing queue, and its later processing will result in identification of the correct grand total and its components.

This extended iterative attempt at checksum validation is ineffective for documents that report zero income. Because there are significant numbers of such documents in our target data set, the tool attempts to identify this special category of documents when the following conservative conditions are satisfied:

- The number is a single zero;
- The number occurs on a line containing the phrase for “Total” (合計);
- The phrase for “Total” has not previously been encountered in the document; and
- There are more single zeros than other values among the preceding numbers in the current **Assumption**, excluding those excluded by the penalty mechanism.

Comparison of the results of this heuristic against original documents has shown this to be a highly reliable indicator that the document does indeed report zero income. The program flow for processing **Assumption** objects is illustrated in Figure 3. For a more complete description of the logic, interested readers may wish to refer to the **PursueTotals** class in the source code of the tool.

6 Conclusion

As an initial trial of the tool during the preparation of this paper, PDF files containing the 2205 most recent financial statements filed by national-level Japanese NPO non-profits were downloaded and processed. This returned 863 validated totals (39.1% of the total document set) with elements of income classified by category. Zero income was identified with confidence in a further 194 documents (8.8% of the set), for a total overall recognition rate of 47.9%. For our statistical purposes, this is more than adequate.

Three important lessons emerged from this initial trial. First, it appears that the logic applied to statements in the *double-entry* format described above failed comprehensively, and must be revisited. The assumption that all lines in such statements consist of three numbers was mistaken. By convention, columns in which a number is zero is often left blank. This gives rise to a further layer of ambiguity that must be addressed when performing this style of line validation.

Many resolution failures resulted from OCR corruption of the digits, an error which no post-processing tool will be able to remedy. The OCR training data used for the trial can and should be improved by extracting character images

from the set of failed documents. This can be expected to substantially improve the recognition rate against the specific document set targeted by our research. Such improvements in the OCR layer are, of course, specific to our particular to our particular project and document set.

Because the categorization of income is based on pattern matching of the “hinting” phrases in the OCR output, there are occasional errors in income categorization. Addressing these for ultimate publication of our data sets will require human intervention. However, because the numbers and totals in the validated returns are known to be correct, this final proofreading can be carried out quickly by human operators with minimal training, given a software interface designed for this specific purpose. Preparation of the necessary proofreading infrastructure is planned for the next phase of development.

This paper has discussed an application of open source OCR technology to financial records stored as graphic data. Techniques for recovering numeric structures and essential metadata from heterogenous documents have been outlined and tested. The results of an initial trial indicate that this method is useful as a means of recovering a significant proportion of such records for the purpose of statistical analysis. While the development of this tool has been driven by a specific social science research project, the code is public, the performance of the tool is proven, and the approach should prove useful as a means of recovering such data in many other areas, including discovery proceedings in the context of litigation.

References

1. F. Bennett & J. Whitney, “Activism in the Wild: A Preliminary Study of NPO Incomes,” 法政論集 [Hōsei Ronshū] (to appear)
2. 東京都生活文化スポーツ局 [Metropolitan Tokyo, Bureau of Citizens, Culture and Sports], “NPO 法人情報公開 [NPO Information Disclosure]”, Mar. 26 2008; <http://www.seikatubunka.metro.tokyo.jp/index4files/etsuran.htm>
3. S.V. Rice, F.R. Jenkins & T.A. Parker, “Fourth Annual Test of OCR Accuracy,” Information Science Research Institute, University of Nevada, Las Vegas, 1995; <http://www.isri.unlv.edu/downloads/AT-1995.pdf>
4. R. Smith, “ReadMe: Important Information All Tesseract Users Need to Know,” Oct. 11, 2007; <http://code.google.com/p/tesseract-ocr/wiki/ReadMe>
5. N. Willis, “Google’s Tesseract OCR Engine is a Quantum Leap Forward,” *Linux.com*, Sep. 28, 2006; <http://www.linux.com/articles/57222>
6. 特定非営利活動促進法 [Act for the Promotion of Non-Profit Activities], art. 1 (Law no. 7, Mar. 25, 1998).
7. 一般社団法人及び一般財団法人に関する法律 [Act Concerning Ordinary Civil Corporations and Ordinary Civil Foundations] (Law no. 48, 2006)
8. 公益社団法人及び公益財団法人の認定等に関する法律 [Act Concerning the Certification of Public Benefit Civil Corporations and Public Benefit Civil Foundations] (Law no. 49, 1998).
9. Post by Ray Smith, 10:53am, Jan. 19, 2008; <http://groups.google.com/group/tesseract-ocr/> (“[W]e are actively developing (and using) Tesseract here at Google, and we are committed to putting our improvements back into the open source code-base.”)

9

Conceptual Search – ESI, Litigation and the issue of Language

David T. Chaplin
Kroll Ontrack

Across the globe, legal, business and technical practitioners charged with managing information are continually challenged by rapid-fire evolution and growth in the legal and technology fields. In the United States, new compliance requirements, amendments to the Federal Rules of Civil Procedure (FRCP) and corresponding case law, along with technical advances, have made litigation support one of the most exciting professions in the legal arena. In the UK, revisions to the Practice Direction to CPR Rule 31 require parties in civil litigation to consider the impacts associated with electronic documents.

One emerging technology trends—both aiding and complicating the management of electronically stored information (ESI) in litigation in the US, EU and UK alike—is the notion of “conceptual search.” This paper focuses on the evolution of conceptual search technology, and predictions of where this science will take legal professionals and technical information managers in coming years and a look at the advantages conceptual search can provide in dealing with the issue of language.

This paper will focus primarily on the latent semantic analysis approach to conceptual search and why this approach is advantageous when searching ESI regardless of the language used in the documents, even to the extent of allowing for cross language searching and accurate searching of documents that contain co-mingle foreign terms with the native language. In order to discuss the language issue the following topics will first be established;

- What is conceptual search?
- Dominant approaches to conceptual search.
- Conceptual Search as a Strategic Litigation Tool

What Is Conceptual Search?

Conceptual search was born out of a need to better locate information in the context of a changing corporate language. Legal teams require access to the information they need to make better, more informed decisions about their cases. Not only is the amount of information growing, there are also significantly more terms being used within the normal corporate lexicon. Abbreviations, acronyms, text and email slang, along with industry and corporate specific terminology, are continually progressing. It is becoming increasingly important that search technology adapts to the changing use of language and the ever-growing amount of information.

Conceptual search is defined as the ability to retrieve relevant information without requiring the occurrence of the search terms in the retrieved documents. Most search technology in use today is traditional keyword search that requires the search term to appear in the retrieved documents. Many of these traditional search engines have

mimicked conceptual search through the use of synonym lists and other human-maintained query expansion approaches. True conceptual search retrieves relevant information in a way that does not require the presence of the search terms without the use of query expansion or independently maintained lexicons, taxonomies or synonym lists. This is why conceptual search is distinctly different from keyword search and is the key to why it is able to adapt to changes in language and the use of slang. Conceptual search allows you to locate information about a topic by understanding what words mean in a given context.

The conceptual search engine must measure subtle patterns and relationships that occur in language. The importance of understanding the context of information is amplified when you consider the complexity of language. Effective search requires the search engine to address synonymy (different words with same meaning) and polysemy (same word with different meanings). For example, cellular means something different when the context is biology versus wireless communications. Conceptual search understands these differences and, in effect, smoothes out the idiosyncrasies of speech by analyzing words and how they are used in context. The measurement of how terms are used in context provides the conceptual search engine with the ability to learn new terminology without human intervention.

Dominant Approaches to Conceptual Search

There are two basic approaches to conceptual search: statistical and linguistic. Statistical methods usually learn from text and do not require any pre-built language models. Statistical methods analyze how terms are used within the document collection to be searched. The statistical method determines the underlying structure of the language based on the documents in the collection. Linguistic methods, including natural language processing (NLP) and syntactic approaches, require models of language that are created and maintained by humans. These models are based on insight into the language and content, or from a training set of related text in order to find universal properties of language and to account for the language's development.

There are also two basic methods in producing conceptual search: automatic and manual. Automatic methods allow you to present any source of information to the system without considering structure or syntax. The automatic method allows for the engine to learn as a new language is introduced to the document collection without any human intervention. Manual methods require humans to create and maintain a taxonomy, ontology or synonym list in order to create and maintain relationships. The knowledge is fixed and will have to be altered to account for new vocabulary or relationships.

One may further classify conceptual search by which scientific method of learning is applied. Again, there are two basic approaches: supervised and unsupervised. Supervised learning requires feedback to improve and to initially specify what needs to be learned. Explicit examples need to be supplied to the system for the engine to

learn. Unsupervised learning is fully autonomous and can arrive at an optimal solution without requiring user feedback or pre-defined training sets.

Finally, conceptual search technology can be query and non-query based. Methods have been developed that enable conceptual search technologies to automatically cluster or folder documents that are similar in theme. These clusters are labeled and provide the business user with the ability to navigate a large set of information that is organized and appropriately labeled without having to issue a query. The ever-increasing influx of information into critical knowledge management systems requires improved methods automatically organizing and making available documents, without requiring the user to know what search to perform. This approach can also be used to enhance or refine existing corporate taxonomies or to provide a “snapshot” of large document collections.

Business professionals, attorneys and litigation support professionals are spending more and more time searching for information to make better and faster decisions. Conceptual search reduces the number of queries, results sets and redundant hits in the standard process of collecting, reviewing and producing documents in discovery. Ultimately, conceptual searching techniques allow legal teams to retrieve the maximum number of relevant documents, including information that would not ordinarily be found through keyword searches.

Conceptual search also simplifies the process. The legal team enters a phrase or sentence and the technology organizes corresponding documents into groups of topics and sub-topics available for document review. For example, if a reviewer knows that all documents in a particular folder are related to stock options and all documents in another folder are related to going out to lunch or birthday celebrations within an office, the reviewer will be able to move through the documents with the level of speed and precision needed to make the most efficient decision about whether the document is important to the matter at hand.

Further, conceptual search provides an intelligent information access layer that sits between the data and the person conducting the search. The value of this technology is important because it provides:

- ***Contextual location of data:*** Relevant information is retrieved based on context, resulting in better and more informed decisions.
- ***Faster identification of data:*** A more advanced understanding of the information is achieved, facilitating faster location of relevant information via better, more accurate search results, which provide quicker decision-making ability.
- ***No other technology needed:*** The engine does not require a query language, providing a faster path to productivity with no training required.

- ***Automated application of the technology:*** An intelligent layer is created that understands your information and continues to learn, providing the ability to automate decisions without human intervention.
- ***The ability to learn more as data volumes increase:*** An intelligent layer that “learns” sits between the business professional and the critical business information, providing accurate and relevant search results as language and terminology change and shift.

Simply stated, conceptual search is the key technology that can facilitate better and faster business decisions in a knowledge economy. Conceptual search provides a mechanism to deliver the right information to the right person at the right time.

Concept search has been available for several years as a tool to help legal and business professionals review data that has already been collected. Concept search can also be used before the document review and production begins for strategic analysis, witness identification, early fact assessment and search term formulation.

Conceptual Search as a Strategic Litigation Tool

When a new investigation or lawsuit begins, US and UK lawyers must start the process of trying to answer the who, what, where, when, why and how questions. Sometimes lawyers have a reasonably good understanding of the people, places and things early in the case, but other times they do not. Rarely, however, will the lawyer possess that knowledge to the degree that allows full early case assessment and a full understanding of who the potential witnesses are and what happened in the case.

Concept search can dramatically improve the speed at which the lawyers develop their case theories, increase the accuracy of the analysis, and decrease the expense of the process. Concept search can help attorneys identify people involved in the dispute, sift through mountains of data and provide an objective, machine-generated group of data with similar context and improve the accuracy of the typical “search-term” approach to data analysis.

Starting with even a limited amount of information about the case, the attorney will be able to identify one or two witnesses who may have knowledge of relevant facts. Through the use of concept search technology, names of other potential witnesses may be dropped into search groupings without requiring the use of search terms, without knowing in advance the names of these individuals, without having to account for misspellings or abbreviations and without having to look at the “to” or “from” lines in email headers. Armed with this information at the beginning of a case, attorneys should more quickly focus on the most important witnesses, even those who are not part of the organization, such as customers, suppliers, competitors and potential wrongdoers.

In addition, having earlier witness identification information will help the legal team ensure that they have preserved data for the right group of custodians. Rather than having to start the data identification process by interviewing each person or by preserving “everything,” early use of concept search can help the legal team hone in on who is a potentially important witness. The concept search results can then be fine-tuned with custodian interview and analysis to ensure the preservation plan is complete.

In the early investigation phase of a case, lawyers frequently know very little about the facts. The investigation may start with nothing more than an anonymous call to an ethics hotline or an allegation of potential wrongdoing by a single employee. Through the use of concept search, the legal team can analyze the data of the accused wrongdoer and quickly profile the subject matter of the data. As the legal team rapidly culls out the irrelevant information, the potential facts become clearer and other witnesses emerge as possible subjects of the investigation. In incremental steps, the legal team can then collect data of others and run concept search technology against that data for sorting and grouping. This technology will help the team get a picture of the facts more quickly and more cost-effectively than the typical method of having a team of lawyers plod through every email or try to formulate guesses at search terms to zero in on the issues.

For years, lawyers have tried to develop the perfect set of search terms, the unobtainable objective of which is to find all relevant data while excluding all irrelevant data. Taking an overly narrow approach to search terms results in the team missing relevant data, but taking an overly broad approach will leave the legal team with much more data than it needs to review.

Lawyers spend hours and hours making, refining and fighting about search term lists. Typically, lawyers for the producing party want a small, narrow list, but lawyers for the requesting party want a large, broad list. But no human being is capable of developing a search-terms list that factors into account the taxonomy and lexicon of the data, nor can any human anticipate all of the abbreviations, misspellings, or “code” language intended to deceive that are prevalent in the data. Concept search can help. It has been repeatedly shown that two people will use the same term to describe something less than 20% of the time. In information retrieval this phenomenon is called term mismatch. The impact of term mismatch is amplified when you factor in that most search engine queries are short and many are a single term. Conceptual search smoothes out this issue by analyzing the context of the terms in the corpus of text being searched.

We may be years away from the time that courts and litigations on either side of the Atlantic Ocean use concept search in lieu of search terms to identify relevant data. But concept search can be used today to fine-tune the search term approach to data identification that litigants are comfortable using.

Concept search can be used to group the data before search terms are developed. The grouped data could be reviewed by the producing party and used to develop search terms to propose to the requesting party. The requesting party, on the other hand, could apply concept search technology to a set of production data that had been identified solely by the use of search terms. Analyzing the grouped data, the requesting party could then provide the producing party with additional search terms to apply against the main data collection. Approaching term-based data productions iteratively has always been the most accurate approach. Including concept search technology in this iterative approach makes the process even better.

The issue of Language

Latent Semantic Analysis (LSA) is a statistical approach to information retrieval that is designed to analyze how terms are used in context and measures the correlation between all the terms in the corpus of text being searched. This means that each term is in fact a token and hence the language of the term is irrelevant. What is relevant is how that term / token is used in context with all the terms / tokens in the corpus of text. LSA by its very approach to text analysis and retrieval is language independent and has the ability to learn the relationships between terms in an automatic an unsupervised indexing scheme. Proper parsing and tokenizing of the language (especially in the case of double byte languages such as Chinese and Japanese) is required and the need exists for a well thought out stop word list telling the search engine what terms should not be indexed due to the noise they would create.

LSA does not have any need to analyze parts of speech or sentence structure which natural language processing requires and in so doing makes the statistical approach a better information retrieval solution. When multiple languages are being processed or when cross lingual or multi-lingual documents are present the ability to understand relationships between terms is critical. LSA with its ability to measure the correlation between terms assists information retrieval in environments containing documents with acronyms, abbreviations, slang from the integration of chat like communications within corporate emails, multi-lingual text and documents introducing new and expanding terminology. In litigation events these conditions exist and they present challenges in processing the ESI and properly preparing for the litigation.

Information retrieval challenges in litigation within the European Union are amplified by the numerous languages present in the union with twenty-seven independent states sharing common business interests. Conditions exist that heighten the probability of many search challenges due to language. The critical nature of processing and making available for search ESI in a litigation requires careful consideration of the tools that will be utilized.

The Future for Conceptual Search

One thing is clear, the use of conceptual search and document clustering technologies have been utilized in the litigation process before case law and legal opinions have

called for the utilization of advanced search solutions. In the United States this has definitely been the case and the same environment exists around the world. The volume of email and other ESI is a consistent problem regardless of where the litigation is taking place. Legal practitioners will always react to the issues of e-discovery in different ways but will always be a segment that will attempt to get ahead of the problems by using new technologies including advanced search tools.

While many issues in discovery are the same in the US, UK and EU the application of advanced search will need to accommodate differences in the collection and review process, regulations and data protection concerns and the growing likelihood that the litigation will require processing data from different countries encompassing many languages. For instance, the EU countries have data protection laws (Council Directive 95/46/EC, 1995 O.J. (L. 281)31 (EC)) that are drastically different from the US in regards to what is considered personal data and broadly defining processing as including collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction. While search is search and discovery is still discovery, how and when you are able to utilize the advanced tools will differ based upon these regulations. The US does not have the same data protection hurdles to discovery with the courts not receptive to most data protection arguments.

In the UK, The Practice Direction to CPR Rule 31 states that parties should consider electronic documents in a litigation. Lawyers are using technology as a mean to explore and better manage electronic disclosure. Savvy lawyers are positioning themselves as expert in electronic disclosure by exploring advanced discovery tools in the period of time prior to the existence of case law. Conceptual search and document clustering technologies are beginning to be implemented as the UK legal community embraces the challenges associated with managing electronic evidence. As in the US the need to focus data collection and reduce the data lawyers need to review in the initial stages of a case is critical. The reduction of data in the early stages is effective in decreasing the time and cost of processing and reviewing the information. Over time, the integration of advanced tools deeper in the litigation process will improve the discovery task as lawyers learn how to apply technology to a problem that is created by technology.

US, UK and EU litigators and business professionals alike are increasingly relying upon technology, like conceptual search, to do their jobs. As more business and legal professionals collect and exchange ESI for multilingual business, litigation, and regulatory purposes, search technology will continue to improve. No matter the global location, one tenet rings true -- the days of searching through file cabinets to locate information are gone. Instead, search technology has and will continue to become an integral part of the corporate and legal business culture in locating, preserving and exchanging electronically stored information.

10

Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings

CaseMap issue linking in UK civil proceedings

By Chris Dale of the E-Disclosure Information Project

Introduction

There is an increasing focus in UK litigation on early identification of the issues between the parties in the form of a Court-approved List of Issues on which all other stages from pleadings through to trial must rely, and on which the facts, the documents and every other element of the case must be hung.

CaseMap is a well-established software application owned by LexisNexis whose primary function is to facilitate the identification of the key components in a case – the facts, people, documents etc – and to help make sense of how these interact with the issues and other elements of the case.

This paper explores briefly how the functionality of CaseMap matches the issues-based focus of the UK courts at all levels.

Focus on the Issues

There is, of course, nothing new in the centrality of the issues in litigation. They are the labels given to the matters in dispute. Every stage from the pleaded case through to final judgment is about the issues, and the facts which underlie them and which must be proved or disproved. The rules and procedures all depend upon, and aim towards, identifying, presenting, and debating, the issues.

The most significant recent development in case management is the **Report and Recommendations of the Commercial Court Long Trials Working Party**¹, released in December 2007 and presently undergoing a trial period in the Commercial Court. Despite its name, its principles are not confined either to long trials or to the Commercial Court. It involves no new law nor variations to the Civil Procedure Rules, and whilst other courts lack the formal structure of a managed trial period, judges are at liberty under their inherent powers of case management to adopt such of the Report's practices as are proportionate to the cases before them.

The Working Party decided² that a “new style, judicially settled, List of Issues” should become “the keystone to the proper management of all Commercial Court cases”. This List of Issues “should be the key working document in all Commercial Court cases, whether small or large and whether involving few or many issues”. It should be a Court Document and “should, once settled, be the basis on which decisions are made about the breadth and depth of disclosure, provision of witness statements, what experts will be permitted and, ultimately, the shape of any trial.” The List of Issues should be “structured and sub-divided”.

Using the List of Issues

Having thus established the importance of the List of Issues, the Recommendations go on to relate almost every other aspect of their management suggestions to that list. It urges a “surgical” approach to disclosure³ and goes on:

¹ http://www.judiciary.gov.uk/docs/rep_comm_wrkg_party_long_trials.pdf

² Para 51 of Commercial Court Recommendations

³ Para 60 *ibid*

“this power to make more specific orders for disclosure, if necessary issue by issue, must be utilised more often by the court. This will be done using the List of Issues, discussed above. Using a more specific approach will also entail the use of a new type of document, a disclosure schedule which will be, effectively, a “shopping list” for disclosure⁴. “

Appendix 3 to the Recommendations is such a “shopping list” in the form of a Disclosure Schedule which shows the Issues, and what each party says about the Disclosure needed in respect of the issues, and with a column to show the order made in respect of each issue.

There are similar provisions as to witness statements which

“...must identify, by reference to the List of Issues, the particular issues on which that witness is giving evidence. This can best be done by having appropriately worded headings in the witness statement”⁵.

There is provision for solid practicalities:

“Where disclosure has been given electronically and it is possible to include a hyperlink to documents referred to within the witness statement, this should be done”⁶.

Expert evidence is to be handled in a similar way:

“The List of Issues should identify, in summary form, the issues on which expert evidence is required, and permission should be limited to expert evidence in relation to those issues. These expert issues may be identified when the List of Issues is first settled or subsequently.”⁷

Other initiatives

In parallel with the above, HHJ Simon Brown QC, a Designated Mercantile Judge in Birmingham, is promoting the application of similar principles in his court.

“What I want to know, is this: what is the case about? Which of the pleaded issues really matter in getting to the heart of the dispute? Can we split the case up and limit disclosure to the subjects which matter, or which matter most?”⁸

Judge Brown has also pioneered a form of draft standard directions order which is premised in large part on the narrowing of the material before the court to the issues identified as central to the case.

The Application of CaseMap to judicial and procedural initiatives

CaseMap is a central case memory for critical case knowledge that can be used to organise information about the key facts, documents, case of characters, issues and case law in every matter. CaseMap makes it easy to evaluate these case details and then to communicate this information to clients, colleagues and the court.

⁴ Para 61 *ibid*

⁵ Para 71 *ibid*

⁶ Para 75 *b ibid*

⁷ Para 83 *b ibid*

⁸ Conference speech January 2008

CaseMap helps lawyers make sense of their cases by facilitating identification of the central issues and by providing a mechanism to allow practitioners to focus tightly on any particular aspect.

CaseMap treats all the elements of a case as “objects” which can be linked together - persons, organisations, documents, physical evidence, events, places, pleadings etc are all “objects”. It also provides for Issues to be cross-linked with each of these object-types and Link Summary (LS) fields make it easy to display reports based on relationships between different types of case information.

Linking the various elements of a case together in this way helps CaseMap users make sense of a case by providing a way to organise case information, evaluate the relationships between the various aspects of the case, and then communicate case knowledge to colleagues, clients and the Court.

A key feature of CaseMap is the Fact-by-Issue report which allows litigators to create case summaries focussed on the issues in the case.

One of CaseMap's particular strengths is the “send to CaseMap” feature found in most mainstream litigation and other applications, notably Adobe Acrobat. Whole sets of records can be selected from a document database and mapped to CaseMap document fields. Passages can be extracted from documents and sent to a CaseMap Facts spreadsheet. The new Fact record links back to the passage in the document, so that a single click brings up the documentary source of the fact.

Other linked objects from passages in pleadings, to the people referred to in them, to the documents in which they are mentioned, can similarly be linked to and reported upon. The case analysis features in CaseMap are supported by related tools, including TimeMap which provides a graphical display of timelines from selected data-based information.

These concepts – the ability to send a sub-set of documents or text extracts to CaseMap and the way in which all the facts and issues can be interlinked (and thus followed whichever of them is the starting-point), map well to the case management regime described above. A core data set can be established at the outset and supplemented as new facts, dates, players and documents are brought into play. This can fluctuate with Issues lists which inevitably change as time passes.

It ties well with Judge Brown's insistence that parties focus on “which of the pleaded issues really matter” and on the facts which must be proved or challenged in respect of those issues. At a time when witness statements are under attack for prolixity and lack of focus (as they are) the CaseMap model helps impose a structure, particularly as the objects common to more than one witness can be re-used between multiple witnesses and in respect of each issue to which they pertain.

A senior US litigator put it to me in this way: the mere act of assembling the core objects in CaseMap and creating the links helps to make sense of them, with the bonus that the result is available for others in the team to share and supplement. The shared access may extend beyond the legal team and out to the experts.

There is an under-estimated benefit which follows from standardisation on CaseMap. Although each case will necessarily be different and may use different objects, the overall form is identical for every case, with the same menus and tables. Consistency aids sensemaking. A user can pick up a dormant case and recall its peculiar facts and issues instantly. Similarly, a supervising partner can keep an eye on multiple cases, particularly as the data includes unresolved questions and unfinished tasks.

CaseMap's main use in the UK market has hitherto been largely for criminal cases. The new focus in the civil courts on the Issues List makes these functions increasingly relevant to civil cases.

Summary

This brief note cannot do more than refer to the procedural developments in the courts. It does even less justice to the wide range of functions and features available in CaseMap. This is enough, however, to indicate that the emphasis on a List of Issues, and on the facts, the documents, the evidence of witnesses and the opinions of experts which are linked to those issues, map extremely well to the purpose for which CaseMap is intended and which it performs very well.

There is a further point. The fact that the heavily issues-based procedure can be applied in any court means that firms of all sizes and users of all skill levels will have to grapple with it. CaseMap provides a simple, powerful, cost effective way to help litigators make sense of their cases on an issue by issue basis. . The range of its possible uses, and the scope of its functions, is extremely wide, but the core features are easily learned. CaseMap has the potential to introduce electronic data handling to non-experts at a time when litigation at all levels must necessarily be run electronically and economically.

Chris Dale
Oxford
14 May 2008

T: 01865 463033 M: 07770 580640 E: chrisdale@chrisdalelawyersupport.co.uk

Supplemental Reading

11

Jason R. Baron, Esq.
Director of Litigation
Office of General Counsel
U.S. National Archives and Records Administration
College Park, Maryland
jason.baron@nara.gov

Compilation of Selected Recent U.S. Case Law & Commentary Referencing Search & Information Retrieval Methods
Updated as of June 15, 2008

I. *Post-Dec 2006 Federal Rules of Civil Procedure Changes*

A. *Cases*

Ameriwood Industries, Inc. v. Liberman, 2007 WL 685623 (E.D. Mo.) (court orders expert report with number of “hits” based on negotiated search terms, with expectation that parties will continue to meet and confer to refine search based on false positives)

ClearOne Communications, Inc. v. Chiang, 2008 WL 920336 (D. Utah) (court adjudicates dispute over conjunctive versus disjunctive operators between search terms)

Disability Rights Council of Greater Washington, et al. v. Washington Metropolitan Transit Authority, 242 F.R.D. 139 (D.D.C. 2007) (Facciola, J.) (proposes use of concept searching as possible supplement to keyword searches)

Equity Analytics, LLC v. Lundin, 248 F.R.D. 331 (D.D.C. 2008) (Facciola, J.) (citing to *U.S. v. O’Keefe*, court questions effectiveness of keyword terms used in search conducted)

Haka v. Lincoln County, 246 F.R.D. 577 (W.D. Wis. 2007) (where parties were initially unable to agree on scope of search terms to be used against four terabytes of data, and where costs of search were on par with amount of damages at stake, court ordered parties to divide cost of a search using a narrowed set of terms, but that defendant-public sector entity would pay 100% of the cost of any subsequent relevance and privilege review)

United States v. O’Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008) (Facciola, J.) (in criminal case, court orders further explanation of whether keyword searches were thorough, citing to authorities arising in civil case law, and suggesting that in light of interplay of the sciences of computer technology, statistics, and linguistics, expert testimony may be needed in this complex area)

Qualcomm Inc. v. Broadcom Corp., 2007 WL 2296441, at *33 (S.D. Cal.) (sanctions opinion involving underlying failure to disclose 200,000 emails prior to trial, where court

found “incredible that Qualcomm never conducted such an obvious search” using certain keywords).

Victor Stanley, Inc. v. Creative Pipe, Inc., 2008 WL 2221841 (D. Md.) (Grimm, J.) (party deemed to have waived attorney-client privilege in failing to sustain their burden that the keyword search method used to identify privileged documents before turning documents over to opposing counsel was reasonable under the circumstances).

Williams v. Taser Intern, Inc., 2007 WL 1630875 (N.D. Ga.) (court adjudicates search protocol with keywords plus use of simple Boolean operators)

B. Law Reviews, Commentaries, and Miscellaneous Publications

Jason R. Baron, *The TREC Legal Track: Origins and Reflections on the First Year*, 8 Sedona Conference Journal 251 (2007) (available on WESTLAW and LEXIS)

Jason R. Baron, Douglas W. Oard, David D. Lewis, *TREC-2006 Legal Track Overview*, http://trec.nist.gov/pubs/trec15/t15_proceedings.html (item 4)

H. Christopher Boehning and Daniel J. Toal, “In Search of Better E-Discovery Methods,” *New York Law Journal*, April 23, 2008, available at <http://www.law.com/jsp/legaltechnology/pubArticleLT.jsp?id=900005509469>.

David Fishel, *Defending the Accuracy of Phonetic Audio Search in Civil Discovery*, (Nexidia), available at <http://www.umiacs.umd.edu/~oard/desi-ws/> (DESI Workshop listed below)

Jeffrey Gross, *Comparing the Utility of Keyword and Concept Searches*, Digital Discovery & E-Evidence, Vol. 7, No. 9, (Sept. 1, 2007) (available online)

(Hon.) Ronald Hedges, *Rule 702 and Discovery of Electronically Stored Information*, Digital Discovery & E-Evidence, Vol. 8, No. 5 (May 1, 2008) (discussing *U.S. v. O’Keefe*)

Mia Mazza, Emmalena K. Quesada, & Ashley L. Stenberg, *In Pursuit of FRCP I: Creative Approaches to Cutting and Shifting Costs of Discovery of Electronically Stored Information*, 13 RICH. J.L. & TECH. 11 (2007), <http://law.richmond.edu/jolt/v13i3/article11.pdf>. (concept searching)

George L. Paul and Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10 (2007), <http://law.richmond.edu/jolt/v13i3/article10.pdf>. (concept searching)

Mark V. Reichenbach, *In Support of Concept Search and Content Analysis*, <http://www.metalincs.com/resources/> (Metalincs white paper)

Herbert L. Roitblat, *Search and Information Retrieval Science*, 8 Sedona Conference Journal (2007) (available on WESTLAW and LEXIS)

Sedona Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery (August 2007 public draft),
http://www.thesedonaconference.org/content/miscFiles/publications_html

Sedona Principles, Second Edition: Best Practices Recommendations & Principles for Addressing Electronic Document Production (June 2007) (Principle 11 discusses search methods),
http://www.thesedonaconference.org/content/miscFiles/publications_html

Stephen Tomlinson, Douglas W. Oard, Jason R. Baron, Paul Thompson, "Overview of the TREC 2007 Legal Track," *available at*
http://trec.nist.gov/pubs/trec16/t16_proceedings.html

C. Ongoing Research and Workshops

<http://trec-legal.umiacs.umd.edu/> (NIST TREC Legal Track)
(see also "Open Letter to Law Firms and Companies in the Legal Tech Sector Re: Invitation to Participate in the TREC Legal Track," at <http://trec-legal.umiacs.umd.edu/> and at www.thesedonaconference.org/publications.)

<http://www.umiacs.umd.edu/~oard/desi-ws/> Workshop on Supporting Search and Sensemaking For Electronically Stored Information in Discovery Proceedings (DESI), Eleventh International Conference on Artificial Intelligence and Law, Palo Alto, June 4, 2007

<http://www.cs.ucl.ac.uk/staff/S.Attfield/desi/> Second International Workshop on Sensemaking for Electronically Stored Information in Discovery Proceedings (DESI II), London, U.K., June 25, 2008

<http://www.ediscoveryinstitute.org/> (legal nonprofit research project)

D. Blogs, Webinars, etc.

<http://ralphlosey.wordpress.com/2008/06/08/hundredth-blog-thoughts-on-search-and-victor-stanley-inc-v-creative-pipe-inc/>

http://abajournal.com/news/e_discovery_disclosure_goof_waived_attorney_client_privilege_judge_rules

<http://ralphlosey.wordpress.com/2007/09/16/sedonas-new-commentary-on-search-and-the-myth-of-the-pharaohs-curse/>

<http://www.ohiotaxlaw.com/legalservices/practice/litigation/ediscotech/eblog/catDisplay.aspx?id=21>

<http://kmpipeline.blogspot.com/2007/09/new-sedona-conference-comments-lend.html>

<http://danmichaluk.wordpress.com/2007/09/22/sedona-conference-search-and-retrieval-draft-paper/>

http://ontolog.cim3.net/cgi-bin/wiki.pl?ConferenceCall_2007_08_23 (J. Baron webinar on “Lawyers, Language, and Legal Risk: Emerging Issues in E-Discovery” for Ontolog Forum, with powerpoints & archived podcast)

<http://chrisdale.wordpress.com/2007/08/31/e-disclosure-2-needles-and-haystacks-3-keywords/>

II. Criminal Case Law Re: Search Protocols

United States v. Adjani, 452 F.3d 1140 (9th Cir. 2006) (issue of the propriety of seizing computers wholesale versus police conducting less intrusive, targeted keyword searching to segregate out intermingled relevant from nonrelevant evidence, discussed in 4th Amendment “search and seizure” context; held, issuance of search warrant arguably overbroad but seizure upheld as reasonable under circumstances)

United States v. Hill, 459 F.3d 966 (9th Cir. 2006) (same)

United States v. Comprehensive Drug Testing, 513 F.2d 1085 (9th Cir. 2008) (same)

12

HU.S. v. O'Keefe
D.D.C., 2008.

United States District Court, District of Columbia.
UNITED STATES of America,
v.
Michael John O'KEEFE, Sr., Sunil Agrawal, Defendants.
Cr. No. 06-249 (PLF/JMF).

Feb. 18, 2008.

Background: In prosecution of Department of State employee for receiving, quid pro quo, gifts and other benefits from his co-defendant for expediting visa requests for employees of co-defendant's company, the United States District Court for the District of Columbia, [Friedman, J.](#), issued an order, [2007 WL 1239204](#), requiring 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. Defendants, who had received the government's submission in compliance with the order, moved to compel, protesting that the government had not fulfilled the responsibilities imposed.

Holdings: The District Court, [John M. Facciola](#), United States Magistrate Judge, held that:

(1) Federal Rules of Civil Procedure may provide guidance in criminal cases in determining whether the production of documents by the government has been in a form or format that is appropriate, and
(2) if electronically-stored information is demanded in criminal case, but the request does not specify a form of production, the responding party must produce the electronically-stored information in the form in which it is ordinarily maintained or in a reasonably usable form or forms.

Order in accordance with opinion.

West Headnotes

[1] Criminal Law 110 ⚡ **627.6(2)**

[110](#) Criminal Law
[110XX](#) Trial
[110XX\(A\)](#) Preliminary Proceedings
[110k627.5](#) Discovery Prior to and Incident to
Trial

[110k627.6](#) Information or Things,
Disclosure of
[110k627.6\(2\)](#) k. Documents or Tangible
Objects. [Most Cited Cases](#)
Federal Rules of Civil Procedure may provide guidance in criminal cases in determining whether the production of documents by the government has been in a form or format that is appropriate. [Fed.Rules Civ.Proc.Rule 34, 28 U.S.C.A.](#)

[2] Criminal Law 110 ⚡ **627.6(2)**

[110](#) Criminal Law
[110XX](#) Trial
[110XX\(A\)](#) Preliminary Proceedings
[110k627.5](#) Discovery Prior to and Incident to
Trial
[110k627.6](#) Information or Things,
Disclosure of
[110k627.6\(2\)](#) k. Documents or Tangible
Objects. [Most Cited Cases](#)
If all of requested documents were produced in criminal case by government in an undifferentiated mass in a large box without file folders or labels, then those documents had not been produced in the manner in which they were ordinarily maintained as required by applicable Federal Rule of Civil Procedure. [Fed.Rules Civ.Proc.Rule 34\(b\)\(2\)\(E\)\(i\), 28 U.S.C.A.](#)

[3] Criminal Law 110 ⚡ **444.20**

[110](#) Criminal Law
[110XVII](#) Evidence
[110XVII\(P\)](#) Documentary Evidence
[110k444](#) Authentication and Foundation
[110k444.20](#) k. Telecommunications. [Most Cited Cases](#)
(Formerly 110k444)
A piece of paper or electronically stored information, without any indication of its creator, source, or custodian may not be authenticated under Federal Rule of Evidence. [Fed.Rules Evid.Rule 901, 28 U.S.C.A.](#)

[4] Constitutional Law 92 ⚡ **4594(8)**

[92](#) Constitutional Law
[92XXVII](#) Due Process
[92XXVII\(H\)](#) Criminal Law
[92XXVII\(H\)4](#) Proceedings and Trial
[92k4592](#) Disclosure and Discovery

[92k4594](#) Evidence

[92k4594\(8\)](#) k. Duty to Preserve. [Most](#)

[Cited Cases](#)

Government's destruction of evidence pursuant to a neutral policy and without any evidence of bad faith does not violate the due process clause if the evidence was destroyed before the defendants raised the possibility that it was exculpatory and the government had no objective reason to believe that it was exculpatory. [U.S.C.A. Const.Amend. 5](#).

[\[5\] Criminal Law 110](#) ➡ [627.6\(1\)](#)

[110](#) Criminal Law

[110XX](#) Trial

[110XX\(A\)](#) Preliminary Proceedings

[110k627.5](#) Discovery Prior to and Incident to

Trial

[110k627.6](#) Information or Things,

Disclosure of

[110k627.6\(1\)](#) k. In General. [Most Cited](#)

[Cases](#)

If electronically-stored information is demanded in criminal case, but the request does not specify a form of production, the responding party must produce the electronically-stored information in the form in which it is ordinarily maintained or in a reasonably usable form or forms. [Fed.Rules Civ.Proc.Rule 34\(b\)\(2\)\(E\)\(ii\), 28 U.S.C.A.](#)

*[15Brenda Jene Johnson](#), Denise Cheung, U.S. Attorney's Office, Washington, DC, for United States of America.

MEMORANDUM OPINION

[JOHN M. FACCIOLA](#), United States Magistrate Judge.

The indictment charges that the defendant, Michael John O'Keefe, Sr., when employed*[16](#) by the Department of State in Canada, received, *quid pro quo*, gifts and other benefits from his co-defendant, Sunil Agrawal, for expediting visa requests for employees of Agrawal's company, STS Jewels.

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\(a\)\(1\)\(E\)](#)).

The first category of "responsive information," as defined by Judge Friedman, was "requests respecting visa applications submitted by or on behalf of STS Jewels employees-including requests for expedited visa interview appointments, decisions granting or denying such interview requests, and the grant or denial of the visas themselves." [Id. at *3](#). This search was to be of the files of the consulates in 1) Toronto, Canada, 2) Ottawa, Canada, 3) Matamoros, Mexico, 4) Mexico City, Mexico, 5) Nogales, Mexico, and 6) Nuevo Laredo, Mexico. [Id.](#)

The second category of "responsive information" was "all written rules, policies, procedures and guidelines regarding the treatment of expedited visa application appointments and visa application approvals at the above-mentioned posts in Canada and Mexico." [Id.](#) The government was also required to "produce any memoranda, letters, e-mails, faxes and other correspondence prepared or received by any consular officers at these posts that reflect either policy or decisions in specific cases with respect to expediting" visa applications. [Id.](#)

As to the latter, Judge Friedman emphasized his expected scope of the search and the necessity for it. He stated:

[I]t now appears from discovery produced on March 21, 2007 that employees below the level of consular officers-including even consulate secretaries and non-U.S. citizen employees-may approve requests for and schedule expedited visa interview appointments. The files of any such persons and the consulates themselves therefore also must be searched. Such communications go directly to the defense of showing that the requests made by or on behalf of STS employees are similar to other requests for expedited visa interview appointments that (it is asserted) have routinely been granted without the provision of anything of value.

[Id.](#)

Defendants, who have received the government's submission in compliance with this Order, have moved to compel, protesting that the government has not fulfilled the responsibilities Judge Friedman imposed. *Memorandum in Support of Defendants' Joint Motion to Compel* ("Def't. Memo") at 1.

I. Detailed Information About the Government's Searches

First, for each location searched, defendants demand a comprehensive description of all of the sources that were searched (both paper and electronic), how each source was searched, and who conducted the search. Deft. Memo at 6 and *Proposed Order*.

In its opposition, the government produced the declaration of Peggy L. Petrovich, the Visa Unit Chief at the United States Consulate General in Toronto, Canada. According to Ms. Petrovich, she, along with her five-member staff, did the *17 following in her effort to comply with Judge Friedman's [April 27, 2007, Order](#):

A. Paper Record Files

1. She [FNI](#) searched “archived hard copies of all Standard Operating Procedures (“SOPs”) to locate Expedited Appointments SOPs dating back to January 2004 that no longer existed in the electronic database.” *Government's Opposition to Defendants' Joint Motion to Compel Discovery* (“Gov.'s Opp.”) at Attachment B, page 2.

[FNI](#). When I use the word “she” I mean Ms. Petrovich and her five-member staff who helped her conduct the search.

2. She searched “archived paper correspondence files to locate expedited appointment requests received via facsimile, correspondence, or electronic mail (email) and the corresponding responses attached to those requests.” *Id.*

3. She searched “hard copy general and chronological files for any other stand-alone documents responsive to the Order.” *Id.*

4. The “search for SOPs yielded archived expedited appointment SOPs covering the period between 2003 through May, 2007.” *Id.* She provided hard copies produced from the electronic sources so that everything she produced to defendants was in the same paper format. *Id.* “The documents printed from the electronic files contain document footers that identify where in the electronic database a document is stored so that it can be located easily.” *Id.*

5. She conducted a “search of the paper correspondence files, the usual and customary storage location for

expedited appointment requests, maintained in three separate five-drawer filing cabinets.” *Id.* This “yielded four drawers with records of expedited appointment requests dating from January 2006 to May 31, 2007.” *Id.* at 2-3. She also searched her own work space and the work spaces of Pat Haye, Jane Boyd, and Althea Brathwaite. *Id.* at 3.

6. “Prior to January 2006, materials relating to expedited appointment requests were attached directly to the non-immigrant visa applications.” *Id.* After one year, the Toronto consulate sends the hard copies of all non-immigrant visa applications to the Kentucky Consular Center for cataloging. *Id.* When the search was conducted, “Toronto only retained [] the hard copies of non-immigrant visa applications received from May 2006 to May 2007” *Id.* All other records had already been shipped to Kentucky. *Id.*

B. Electronic Record Files

1. *Search and Yield*: She searched all active servers and backup tapes (retained for two weeks) and that search yielded “responsive emails, the SOPs previously mentioned, and the NIV (Non-Immigrant Visa) Schedule Calendar located on Toronto's shared public drive.” *Id.*

2. *Parameters of the Search Conducted*: “[T]he electronic search included all email and stand-alone electronic documents, e.g., documents prepared on our office software applications, regarding expedited appointments located on shared drives, personal drives and hard drives for all consular officers and locally-engaged staff, i.e., secretaries and other employees, who approved or scheduled expedited non-immigrant visa interviews, or who played any role in the process.” *Id.* [FN2](#)

[FN2](#). In her declaration, she identified these 19 people by name. *Id.* She also searched electronic depositories identified as “Gold, Toronto,” “Toronto NIV,” “Toronto, Employment NIV Mailbox,” and the files and folders of five former members of the staff. *Id.*

*18 3. *Search Terms*: She used the following search terms: “early or expedite* or appointment or early & interview or expedite* & interview.” *Id.* She had “[t]he Information Management Staff conduct[] the search of personal and hard drives because they have access to all drives from the network server, not just shared drives.” *Id.*

4. *Review of Results*: She reviewed the results of the search and “removed only those clearly about wholly unrelated matters, *e.g.*, emails about staff members' early departures or dentist appointments.” *Id.* She “made sure that all emails residing in the shared email address folders that related to expedited appointments were included in the results ... that were produced in electronic format and provided on cd-rom.” *Id.* at 4.

5. *Deleted Emails*: “According to the Information Management staff, any emails deleted prior to [her] search” in May 2007 are gone. *Id.* Electronically stored information is backed up for two weeks and then the back up tapes are reused and their previous contents obliterated. *Id.* “No other back-up server for electronic documents, either on- or off-site, exists.” *Id.*

6. *O'Keefe Emails*: “All currently existing responsive emails located during the search of Michael O'Keefe's personal drive were included in the cd-rom” that the government gave the defendants. *Id.* Since the hard drives from the computers O'Keefe used were previously seized by the government, they could not be searched. *Id.*

7. *SOPs*: “The only other responsive materials discovered during the electronic search for stand-alone electronic documents were the SOPs [described in paragraph 4, *supra*] and the NIV Schedule Calendar which was provided in hard copy format.” *Id.*

8. *Lack of documents*: There were no responsive documents from “Mike Schimmel, the previous visa unit chief; Peggy Petrovich, the current visa unit chief; Pat Haye, the visa assistant who has main responsibility for processing expedited appointment requests; and, Jane Boyd, the visa assistant who has main responsibility for scheduling appointments for diplomatic and official applicants.” *Id.*

II. Problems with the Government's Production

A. Hard Copy Production

Defendants complain that the government has produced the written documents in a manner which makes it impossible to identify the source or custodian of the document. They point out that they initially requested that the government mark the documents using the familiar Bates system of numbering documents and then advise

the defendants of the Bates range for each individual's responsive documents, *i.e.*, documents from the files of John Doe would be identified in a separate index as “Doe 123-137.” Since the government hasn't done this, defendants demand that the government now produce an index for its entire paper production which shows, for each document, the custodian of the documents, his or her title, the source of the document, whether it is a paper document or electronically stored information, and the Bates number of the document. Deft. Memo at 12 n.8.

1. Use of the Federal Rules of Civil Procedure^{FN3}

^{FN3}. All references to the Federal Rules of Civil Procedure are to the version effective December 1, 2007.

[1] In criminal cases, there is unfortunately no rule to which the courts can look *19 for guidance in determining whether the production of documents by the government has been in a form or format that is appropriate. This may be because the “big paper” case is the exception rather than the rule in criminal cases. Be that as it may, [Rule 34 of the Federal Rules of Civil Procedure](#) speak specifically to the form of production. The Federal Rules of Civil Procedure in their present form are the product of nearly 70 years of use and have been consistently amended by advisory committees consisting of judges, practitioners, and distinguished academics to meet perceived deficiencies. It is foolish to disregard them merely because this is a criminal case, particularly where, as is the case here, it is far better to use these rules than to reinvent the wheel when the production of documents in criminal and civil cases raises the same problems.

2. [Rule 34](#) and the Form of Production of Documents

Under [Rule 34\(b\) of the Federal Rules of Civil Procedure](#), a party, on whom a demand for production of documents has been made, must produce them in the form in which they are ordinarily maintained or must organize and label them to correspond with the categories of the request for production. [Fed.R.Civ.P. 34\(b\)\(2\)\(E\)\(i\)](#). While the Rule is premised on the expectation that the requesting party copies the documents once they have been produced, the more common experience is that the producing party copies the documents for the requesting party, as occurred here.

The Rule was amended in 1980 to prevent the juvenile

practice whereby the producing party purposely rearranged the documents prior to production in order to prevent the requesting party's efficient use of them. [Fed.R.Civ.P. 34](#) advisory committee's note. See [Sparton Corp. v. United States](#), 77 Fed.Cl. 10, 19 (Cl.Ct.2007). In eliminating that practice and requiring the producing party to produce the documents in the same way they were kept, the Advisory Committee intended that there would be equality between the parties in their ability to search the documents. Thus, if the documents were produced as they were kept in the ordinary course of business, the requesting party could not thereafter demand that they be indexed, catalogued, or labeled. [Washington v. Thurgood Marshall Acad.](#), 232 F.R.D. 6, 10 (D.D.C.2005); [Doe v. D.C.](#), 231 F.R.D. 27, 35 (D.D.C.2005). The producing party can, alternatively, label the documents to correspond with the categories in the initial request, irrespective of how the documents were maintained in the ordinary course of their business.

If documents are removed from their original containers and then copied, those copies are not being produced in the manner in which the originals were ordinarily kept, since, in their original condition, the originals were most probably in labeled file folders. Therefore, to reproduce them in the manner in which they were kept would require the producing party to reproduce those file folders and place the appropriate documents in them so that the production replicates the manner in which they were originally kept. If that is not done, federal courts have required the producing party to index the documents to render them usable by the requesting party. See, e.g., [Okla. ex rel Edmonson v. Tysons Food, Inc.](#), No. 05CV329(GKF/SAJ), 2007 U.S. Dist. LEXIS 36308, at *16 (N.D.Okla. May 17, 2007) (requiring producing party to create a "complete and fully accurate index ... showing the box number which responds to each specific Motion to Produce"); [Sparton](#), 77 Fed.Cl. at 16 (criterion is whether the documents are so disorganized that it would be unreasonable for the *20 requesting party to review the documents; producing party may not provide documents in "mass of undifferentiated, unlabeled documents" but must provide them in some "organized, indexed fashion"); [Am. Int'l Specialty Lines Ins. Co. v. NWI-I, Inc.](#), 240 F.R.D. 401, 411 (N.D.Ill.2007) (party providing access to warehoused documents with master index failed to comply with [Rule 34\(a\)](#) where some boxes were inaccurately labeled and 1,778 boxes either had no labels or labels did not provide indicia of contents of boxes); [Wagner v. Dryvit Sys., Inc.](#), 208 F.R.D. 606, 610-11 (D.Neb.2001) (directing plaintiffs to search through volumes of irrelevant information does not comply with

[Rule 34\(a\)](#); that producing party has unwieldy record keeping system that requires much time and effort to find anything is no excuse); [In re: Sulfuric Acid Antitrust Litig.](#), 231 F.R.D. 351, 363 (N.D.Ill.2005) (producing party may not dump massive amounts of documents in no logical order on their opponents; undifferentiated production of everything in boxes will not do). See also [T.N. Taube Corp. v. Marine Midland Mortgage Corp.](#), 136 F.R.D. 449, 456 (W.D.N.C.1991) ("The Court doubts very much whether Defendant complied with the commands of [then] [Rule 34\(b\)](#) that documents be produced as kept in the usual course of business. It is certainly improbable that Marine Midland routinely haphazardly stores documents in a cardboard box.").

[2] I have not seen the documents at issue in this case, but I can say that, if all of the documents have been produced in an undifferentiated mass in a large box without file folders or labels, then these documents have not been produced in the manner in which they were ordinarily maintained as [Rule 34\(b\)\(2\)\(E\)\(i\)](#) requires. To be useful at the consulate, in their original state, they must have been placed in labeled file folders in the file cabinets described by Ms. Petrovich. Without such file folders and labels, it is impossible to understand how anyone who needed to find these documents could have done so.

Defendants have encountered another problem: they have been provided documents from the various consulates in single consulate-specific Bates-numbered series "without providing any information regarding the custodian or source of the documents." Deft. Memo at 6. According to the defendants, this leads them to "guess about the evidentiary value of the documents-i.e., who created a document or on whose computer or in whose file a document was kept." *Id.* Defendants insist that, without knowing the creator of the document and its original location, it will be impossible to authenticate the document and offer it into evidence. *Reply to Government's Opposition to Defendants' Second Motion to Compel* ("Deft. Reply") at 7-8.

[3] A piece of paper or electronically stored information, without any indication of its creator, source, or custodian may not be authenticated under [Federal Rule of Evidence 901](#). There is an obvious solution to this problem, however. I will recommend to Judge Friedman that he deems all documents produced by the government authentic and relieve the government of the task of making the certification required by [Rule 902\(11\) of the Federal Rules of Evidence](#).

3. Relevance

That still leaves the problem of relevance. It may be impossible to ascertain the relevance of a document without any information as to its custodian, source, author or recipient.

Defendants propose the detailed chart of every document that I have described. *Supra* at 6-7. In response, while Ms. *21 Petrovich does not speak to how she organized what was found, the government in its opposition explains that the chart that defendants demand is not necessary because the authors and recipients of the e-mails, correspondence, and memoranda are self-evident from the documents themselves. Gov.'s Opp. at 4. The government insists that forcing consular officials to go back over the hard copy production and create the chart defendants demand would force them to do the entire search process all over again. *Id.*

Since the production has already occurred and the parties

did not discuss this problem before hand, the eggs have been scrambled and the only hope is to try to create a solution that will take into account the needs of the defendants to discern the information they need to prove a document's relevance when it is not immediately evident from the document itself. I will therefore try to create a solution that meets the needs of the defendants to know the author, recipient (if any), date of a document, and in what file it originally was without unnecessarily burdening the government.

First, defendants' counsel and government's counsel shall meet. Defendants will produce for the government's inspection all documents that they claim cannot be identified on their faces by author, recipient (if any), date of creation, and consulate location. The parties will then attempt to arrive in good faith at a stipulation as to the author, recipient (if any), date of each such document, and where the document was found. Defendants will have the obligation to memorialize each stipulation agreed to by the parties by marking it with a Bates number so that there will ultimately be created a joint index as to these documents. Such an index would be formatted as follows:

Bates Number	Author	Author's Title ⁴	Recipient (if any)	Date of Creation	Location of Document
-----------------	--------	--------------------------------	-----------------------	---------------------	-------------------------

FN4. If the role of the author is not obvious from

the author's title, a description of the author's role within the Consulate shall be provided.

JEX 1	John Smith	Assistant to the Visa Unit Chief	Betty Brown	2/2/2005	TorontoCon sulate/Offic e of Betty Brown/File Cabinet/Fol der LabeledExp editedAppoi ntmentCorre spondence
-------	---------------	--	----------------	----------	---

If the government insists that the document is self-identifying and the defendants disagree, the document will have to be put to one side and I will resolve the controversy. I would urge the government to have Ms. Petrovich available by phone and fax since she may be able to look at any document in dispute and promptly provide the identifying information upon which the stipulation may be based.

I intend to consult with the parties as necessary to expedite this process, and I would ask them to call upon me as they are doing this if they believe I can resolve any controversy.

4. Search of Employees' Work Spaces and Other Consulates

Defendants point out that Ms. Petrovich only searched the work space files of four *22 individuals, while she searched the personal electronic files of “24 individuals, as well as other shared electronic files.” Deft. Reply at 5. They also complain that the government has not explained how the search was conducted at other consulates. *Id.*

Both points are well taken. I will therefore direct the government to file a supplemental declaration from Ms. Petrovich as to why she did not search the workspace files of the 24 persons whose electronic files she searched. The government must provide declarations from representatives of the other consulates that were searched indicating how the search was conducted. These declarations should be in the same detail as Ms. Petrovich provided as to Toronto.

B. Electronic Production

Defendants marshal several objections and concerns about the government's search of the electronically stored information. They take the government to task for 1) not interviewing the employees as to their use of electronic means as a form of communication regarding expedited reviews, 2) not having the employees search their own electronically stored information and 3) not indicating what software it used to conduct the search or how it ascertained what search terms it would use. *Id.*

Defendants caution that, if forensic searchware was not used, there is a likelihood that stored e-mail folders in .pst files were either not searched or not searched accurately. They also note that the “government has not said anything regarding the document preservation efforts that were undertaken at the time of the Indictment or at the time the Court issued its April 27, 2008 Order.” I take up these issues in turn.

1. Preservation

[4] The government's destruction of evidence pursuant to a neutral policy and without any evidence of bad faith does not violate the due process clause if the evidence was destroyed before the defendants raised the possibility that it was exculpatory and the government had no objective reason to believe that it was exculpatory. [Arizona v. Youngblood](#), 488 U.S. 51, 57, 109 S.Ct. 333, 102 L.Ed.2d 281 (1988); *In re: Sealed Case*, 99 F.3d 1175, 1178 (D.C.Cir.1996). Accord [United States v.](#)

[Beckstead](#), 500 F.3d 1154, 1158-62 (10th Cir.2007); *Bower v. Quarterman*, 497 F.3d 459, 476-77 (5th Cir.2007) (exculpatory value of destroyed evidence must be apparent before its destruction).

This principle finds its analogue in the Federal Rules of Civil Procedure, which indicate that, absent exceptional circumstances, sanctions will not be awarded for a party's failure “to provide electronically stored information lost as a result of the routine, good-faith operation of an electronic information system.” [Fed.R.Civ.P. 37\(e\)](#).

Defendants protest that there are inexplicable deficiencies in the government's production of electronically stored information, but, as I have indicated in another case, vague notions that there should have been more than what was produced are speculative and are an insufficient premise for judicial action. See [Hubbard v. Potter](#), 247 F.R.D. 27, 30-31 (D.D.C.2008). Accusations that the government purposefully destroyed what they were obliged to produce or knowingly failed to produce what a court ordered are serious. I must therefore remind the defendants of the wise advice given the revolutionary: “If you strike at a king, kill him.” If the defendants intend to charge the government with destroying information that they were obliged to preserve and produce pursuant *23 to Judge Friedman's order or the due process clause itself, they must make that claim directly and support it with an evidentiary basis—not merely surmise that they should have gotten more than they did. If they do not do so within 21 business days of this opinion, I will deem any such claim to have been waived.

2. Metadata

In a footnote, the defendants indicate that “[i]t is not *per se* problematic that the government produced electronic images (PDF or TIF) to the defense instead of native files—as long as the government preserves the original native files and the unaltered metadata associated with those files.” Reply at 7 n.5. Defendants note further that “[i]f metadata becomes important as evidence regarding particular documents, the defense will request that it be provided by identifying such documents by Bates number.” *Id.* Finally, defendants contend that “simply producing native files does not show that those files were produced in the manner in which they were kept in the ordinary course [and that] [t]hat showing must also be made by identifying the custodian and source of the documents produced.” *Id.*

[5] First, Judge Friedman's order does not speak to the format of the production of electronically stored information. Under [Rule 34 of the Federal Rules of Civil Procedure](#), a distinction between documents and electronically stored information is made in terms of the form of production. As established above, a party is obliged to either produce documents as they are kept in the usual course of business or it "must organize and label them to correspond to the categories in the request." [Fed.R.Civ.P. 34\(b\)\(2\)\(E\)\(i\)](#). But if, as occurred here, electronically-stored information is demanded but the request does not specify a form of production, the responding party must produce the electronically-stored information in the form in which it is ordinarily maintained or in a reasonably usable form or forms. [Fed.R.Civ.P. 34\(b\)\(2\)\(E\)\(ii\)](#). Additionally, a party "need not produce the same electronically stored information in more than one form." [Fed.R.Civ.P. 34\(b\)\(2\)\(E\)\(iii\)](#).

If one were to apply these rules to this case, it appears that the government's production of the electronically stored information in PDF or TIFF format would suffice, unless defendants can show that those formats are not "reasonably usable" and that the native format, with accompanying metadata, meet the criteria of "reasonably usable" whereas the PDF or TIFF formats do not.

I appreciate that the government seems ready and willing to produce the documents in native format [FN5](#) and that may obviate the problem. I think it crucial to warn the defendants, however, that they must now secure a stipulation from the government that the electronically stored information will be preserved in its native form with metadata. If the government refuses, defendants will have to move the Court to compel the government to do so, meeting the criteria I have just specified. In the latter event, I expect the government to preserve the electronically stored information in its native format with metadata until the Court rules on the defendants' motion.

[FN5](#). Gov.'s Opp. at 4 n.2.

3. Search Terms and Other Deficiencies

As noted above, defendants protest the search terms the government used. [FN6](#)*24 Whether search terms or "keywords" will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics. See George L. Paul & Jason R. Baron, [Information Inflation: Can the Legal System Adapt?](#), 13

[RICH. J.L. & TECH. 10 \(2007\)](#). Indeed, a special project team of the Working Group on Electronic Discovery of the Sedona Conference is studying that subject and their work indicates how difficult this question is. See *The Sedona Conference, Best Practices Commentary on the Use of Search and Information Retrieval*, 8 THE SEDONA CONF. J. 189 (2008), available at http://www.thesedonaconference.org/content/miscFiles/Best_Practices_Retrieval_Methods_revised_cover_and_preface.pdf. Given this complexity, for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread. This topic is clearly beyond the ken of a layman and requires that any such conclusion be based on evidence that, for example, meets the criteria of [Rule 702 of the Federal Rules of Evidence](#). Accordingly, if defendants are going to contend that the search terms used by the government were insufficient, they will have to specifically so contend in a motion to compel and their contention must be based on evidence that meets the requirements of [Rule 702 of the Federal Rules of Evidence](#).

[FN6](#). Note that the defendants also take the government to task for not interviewing the employees to ascertain how often they used electronic means to create any electronic documents regarding expedited interviews. Reply at 6. But if the search terms used actually captured everything there was to capture, such interviews would be unnecessary.

4. Other Consulates

Judge Friedman's Order requires the government's search to be conducted at other consulates besides the one in Toronto. I assume those searches have been conducted. If they have not, I will direct that the parties appear before me to create a protocol for their search so that the problems the parties have confronted in the search of the Toronto consulate can be avoided.

An Order accompanies this Memorandum Opinion.

D.D.C., 2008.
U.S. v. O'Keefe
537 F.Supp.2d 14

END OF DOCUMENT

13

Victor Stanley, Inc. v. **CreativePipe**, Inc.
D.Md., 2008.

Only the Westlaw citation is currently available.

United States District Court, D. Maryland.

VICTOR STANLEY, INC., Plaintiff

v.

CREATIVEPIPE, INC., et al., Defendant.

Civil Action No. MJG-06-2662.

May 29, 2008.

[Randell C. Ogg](#), Bode and Grenier LLP, [Robert Benjamin Wolinsky](#), Hogan and Hartson LLP, Washington, DC, for Plaintiff.

[James A. Rothschild](#), Anderson, Coe and King LLP, Baltimore, MD, [Frear Stephen Schmid](#), [Frear Stephen Schmid](#) Attorney at Law, San Francisco, CA, [Jeffrey M. Orenstein](#), Goren, Wolff and Orenstein LLC, Rockville, MD, [Joshua J Kaufman](#), Venable LLP, Washington, DC, for Defendants.

MEMORANDUM AND ORDER

[PAUL W. GRIMM](#), United States Magistrate Judge.

*1 The plaintiff, Victor Stanley, Inc. ("VSI" or "Plaintiff") filed a motion seeking a ruling that five categories of electronically stored documents produced by defendants **CreativePipe**, Inc. ("CPI") and Mark and Stephanie Pappas ("M. Pappas", "S. Pappas" or "The Pappasses") (collectively, "Defendants") in October, 2007, are not exempt from discovery because they are within the protection of the attorney-client privilege and work-product doctrine, as claimed by the Defendants. VSI argues that the electronic records at issue, which total 165 documents, are not privileged because their production by Defendants occurred under circumstances that waived any privilege or protected status. Alternatively, as for a subset of nine email communications from M. Pappas to a computer forensics expert Defendants retained to assist them with producing electronically stored information ("ESI"), VSI contends that the attorney-client privilege is inapplicable, and with regard to another two email communications (one draft, the other actually sent) from M. Pappas to one of his attorneys, VSI contends that they are neither privileged nor protected. Finally, as for two email communications

from M. Pappas to two of his attorneys, VSI argues that they are beyond the scope of the attorney-client privilege because they fall within the crime/fraud/tort exception. Defendants acknowledge that they produced all 165 electronic documents at issue to VSI during Rule 34 discovery, but argue that the production was inadvertent, and therefore that privilege/protection has not been waived. As to the various email communications, Defendants argue that they are within the scope of the attorney-client privilege and work-product protection, and that the crime/fraud/tort exception is not applicable. The motion has been fully briefed, Paper Nos. 212, 221, 225, and 230, and I find that a hearing is not necessary. Local Rules of the United States District Court for the District of Maryland, Rule 105.6. For the reasons that follow, I find that all 165 electronic documents are beyond the scope of the attorney-client privilege and work-product protection because assuming, *arguendo*, that they qualified as privileged/protected in the first instance,^{FN1} and assuming further that Defendants properly complied with their obligation to particularize any claims of privilege/protection imposed by [Fed.R.Civ.P. 26\(b\)\(5\)](#), Local Rules of the United States District Court for the District of Maryland, Appendix B, Discovery Guideline 9.c ("Discovery Guideline"), and the orders of this court detailing how such assertions must be demonstrated once they were challenged by VSI,^{FN2} the privilege/protection was waived by the voluntary production of the documents to VSI by Defendants.

Background Facts

The following facts are not subject to dispute. The Defendants' first Rule 34 response was a "paper production," not ESI, made in May 2007. Pl.'s Supp'l Mem. 3, Paper No. 221. Plaintiff objected to its sufficiency, and following a hearing, the court ordered the parties' computer forensic experts to meet and confer in an effort to identify a joint protocol to search and retrieve relevant ESI responsive to Plaintiff's Rule 34 requests. *Id.* This was done and the joint protocol prepared. Pl.'s Supp'l Mem. Ex. 9, Paper No. 221. The protocol contained detailed search and information retrieval instructions, including nearly five pages of keyword/phrase search

terms. It is noteworthy that these search terms were aimed at locating responsive ESI, rather than identifying privileged or work-product protected documents within the population of responsive ESI. After the protocol was used to retrieve responsive ESI, Defendants reviewed it to locate documents that were beyond the scope of discovery because of privilege or work-product protection. Counsel for Defendants had previously notified the court on March 29, 2007, that individualized privilege review of the responsive documents “would delay production unnecessarily and cause undue expense.” Pl.’s Letter of Mar. 29, 2007, Paper No. 79. To address this concern, Defendants gave their computer forensics expert a list of keywords to be used to search and retrieve privileged and protected documents from the population of documents that were to be produced to Plaintiff. *Id.* However, Defendants’ counsel also acknowledged the possibility of inadvertent disclosure of privileged/protected documents, given the volume of documents that were to be produced, and requested that the court approve a “clawback agreement” fashioned to address the concerns noted by this court in [Hopson v. Mayor of Baltimore](#), 232 F.R.D. 228 (D.Md.2005).^{FN3} In response, the court held a telephone conference to discuss the proposed clawback agreement, and thereafter issued a letter order requesting additional briefing by the parties “regarding the burdens associated with conducting a privileged [sic] review of the information to be produced in the time frame required by [the] discovery [schedule] in this case.” Letter Order, Apr. 24, 2007, Paper No. 92. However, on April 27, 2007, Defendants’ counsel notified the court that because Judge Garbis recently had extended the discovery deadline by four months, Defendants would be able to conduct a document-by-document privilege review, thereby making a clawback agreement unnecessary. Defs.’ Letter of Apr. 27, 2007, Paper No. 93. Accordingly, Defendants abandoned their efforts to obtain a clawback agreement and committed to undertaking an individualized document review.

*2 Following their privilege review, Defendants made their ESI production to Plaintiff in September 2007. Pl.’s Supp’l Mem. 5, Paper No. 221. It is noteworthy that by the time of this production, Defendants had discharged their local attorneys, Messrs. Mohr and Ludwig from Meyer, Klipper & Mohr, and brought in

new counsel.^{FN4}

After receiving Defendants’ ESI production in September, 2007, Plaintiff’s counsel began their review of the materials. They soon discovered documents that potentially were privileged or work-product protected and immediately segregated this information and notified counsel for Defendants of its production, following this same procedure each time they identified potentially privileged/protected information. Pl.’s Supp’l Mem. Exs. 11-15, Paper No. 221. Defendants’ Counsel, Mr. Schmid, responded by asserting that the production of any privileged or protected information had been inadvertent. Pl.’s Supp’l Mem. Ex. 17, Paper No. 221. Defendants also belatedly provided Plaintiff with a series of privilege logs, purportedly identifying the documents that had been withheld from production pursuant to [Fed.R.Civ.P. 26\(b\)\(5\)](#). Defs.’ Opp’n Mem. Exs. 4, 6, and 9, Paper No. 225.

The parties disagree substantially in their characterization of how Defendants conducted their review for privileged and protected documents before the ESI productions were made to Plaintiff. Defendants contend that after the joint ESI search protocol was implemented and the responsive ESI identified, their computer forensics expert, Ms. Genevive Turner, “conducted a privilege search using approximately seventy different keyword search terms ... [that] had been decided upon previously by Mr. Pappas, his former attorney, Christopher Mohr, and another attorney, F. Stephen Schmid.... All documents which were returned during the keyword search were segregated and provided to one of Mr. Pappas’ attorneys, John G. Monkman, Jr. for the first phase of the pre-production privilege review.” Defs.’ Opp’n Mem. 4, Ex. 1 (Pappas Aff.) and Ex. 3 (Monkman Aff.), Paper No. 225. This characterization, however, is somewhat misleading. In actuality, after the joint retrieval protocol had been executed, Ms. Turner determined that there were some ESI files (4.9 gigabytes) that were in text-searchable format and others (33.7 gigabytes) that were not. Defs.’ Opp’n Mem. Ex. 2 (Turner Aff. ¶ 7), Paper No. 225. Turner conducted a search for privileged material on the text-searchable files using the seventy keywords developed by M. Pappas, Mohr and Schmid. As to the nontext-searchable files, she

produced them to Monkman for manual privilege review. Turner Aff. ¶¶ 6-7. Monkman reviewed each of the files identified as privileged/protected by Turner based on her keyword searches. Monkman Aff. ¶ 7. Additionally, Monkman and M. Pappas teamed up to begin doing a “page-by-page” manual privilege review of the nontext-searchable ESI files. *Id.* at ¶ 8. According to Monkman: “[t]he second phase of review consisted of page-by-page review of ... [the non text-searchable ESI files], which was undertaken by Mr. Pappas and me. However, due to the compressed schedule and time constraints in reviewing these tens of thousands of documents within the time permitted, this review was undertaken by reviewing the page titles of the documents. Documents whose page titles indicated that the privilege might be applicable were reviewed in their entirety by Mr. Pappas or me. This was the only way for us to complete the unwieldy review of these documents within the time permitted”. *Id.*

*3 The foregoing affidavits create the impression that the keyword search Turner conducted on the text-searchable ESI files, using the seventy keywords developed by M. Pappas and his attorneys, successfully culled out the privileged/protected documents; and this status was confirmed by Monkman's review, and they were withheld from production. Further, the Defendants' characterization of the privilege review suggests that as to the non-text searchable files, Pappas and Monkman did all that could be reasonably expected of them in the time allowed to make the ESI production, which was to review only the title page of the documents and not their entire content. From the affidavits Defendants provided, the court is left to infer that the text-searchable documents that were not flagged by the keyword search Turner conducted were produced to the Plaintiff, as well as the nontext-searchable files that Monkman and M. Pappas determined were not privileged or protected based on their limited title-page review. This is because the Defendants fail to delineate exactly which documents were and were not provided to the Plaintiff, or where the 165 documents at issue were located within the ESI productions made to the Plaintiff.

The implied conclusion that the court is invited to draw, from the limited information provided by the

Defendants, is that the 165 documents that are the subject of the present motion were contained within the population of nontext-searchable ESI files that were produced by the Defendants to the Plaintiff, making their production inadvertent. However, this inference is not so easily drawn.

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. While it is known that M. Pappas (a party) and Mohr and Schmid (attorneys) selected the keywords, nothing is known from the affidavits provided to the court regarding their qualifications for designing a search and information retrieval strategy that could be expected to produce an effective and reliable privilege review. 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. 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. 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. 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. There is no evidence on the record that the Defendants did so in this case. Rather, it appears from the information that they provided to the court that they simply turned over to the Plaintiff all the text-searchable ESI files that were identified by the keyword search Turner performed as non-privileged, as well as the non-text searchable files that Monkman and M. Pappas' limited title page search determined

not to be privileged.

*4 The Plaintiff paints an entirely different picture of the Defendants' privilege review. VSI vigorously disputes Defendants' assertion that the text-searchable ESI received by Defendants' computer forensic expert, Turner, following the execution of the joint search and retrieval protocol was in a format that was difficult to search for privileged or protected materials. Plaintiff contends that it was able to do a keyword search of the text-searchable ESI produced by Defendants in about one hour using a "readily-available desktop search tool." Pl.'s Reply Mem. 3 and Ex. 1 (Slaughenhoupt Aff. ¶ 6), Paper No. 230. VSI further contends that the nontext-searchable files that Monkman and M. Pappas reviewed by looking at the title pages consisted primarily of image files, such as photographs, catalogs, and drawings, which are not likely to contain privileged or protected information. *Id.* at ¶ 9. Most importantly, however, the Plaintiff argues that the Defendants' complaint-that they could not effectively conduct a privilege review of the nontext-searchable files because there were so many of them-is a red herring because "the privileged materials [that are the subject of this motion] were all in text and thus were all searchable using standard text search tools. Contrary to Mr. Pappas' assertion, a majority of the .PDF files in the ESI were searchable using readily available search tools. The ESI contained 9008 .PDF files, the majority of which were searchable and the remaining could have been made searchable using readily available OCR software and/or the native OCR Text Recognition tool within Adobe Acrobat." *Id.* at ¶ 8.

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.

Applicable Law

As this court discussed in some detail in [Hopson, 232 F.R.D. at 235-38](#), courts have taken three different approaches when deciding whether the inadvertent production to an adversary of attorney-client privileged or work-product protected materials constitutes a waiver. Under the most lenient approach there is no waiver because there has not been a knowing and intentional relinquishment of the privilege/protection; under the most strict approach, there is a waiver because once disclosed, there can no longer be any expectation of confidentiality; and under the intermediate one, the court balances a number of factors to determine whether the producing party exercised reasonable care under the circumstances to prevent against disclosure of privileged and protected information, and if so, there is no waiver. *Id.* As also noted in *Hopson*, the Fourth Circuit Court of Appeals has yet to decide which approach it will follow, although individual district courts within the circuit have adopted the intermediate balancing approach. *Id.* at 236 n. 18; see also *Cont'l Cas. Co. v. Under Armour, Inc.*, 537 F.Supp.2d 761, 768 n. 3 (D.Md.2008). As *Hopson* pointed out, however, a careful reading of the Fourth Circuit's decisions regarding waiver of the attorney-client privilege, albeit in contexts not closely related to the facts of this case,^{FN5} suggest that it is more inclined to adopt the strict approach than the intermediate or lenient one. [Hopson, 232 F.R.D. at 236-37](#). Under the strict approach, there is no legitimate doubt that Defendants' production of the 165 asserted privileged/protected documents waived the attorney-client privilege and work-product protection.^{FN6} Even under the intermediate test, however, the result would be the same.^{FN7}

*5 The intermediate test requires the court to balance the following factors to determine whether inadvertent production of attorney-client privileged materials waives the privilege: (1) the reasonableness of the precautions taken to prevent inadvertent disclosure; (2) the number of inadvertent disclosures; (3) the extent of the disclosures; (4) any delay in measures taken to rectify the disclosure; and (5) overriding interests in justice. [McCafferty's, Inc., v. Bank of Glen Burnie](#), 179 F.R.D. 163, 167 (D.Md.1998) (citing cases). The first of these factors militates most strongly in favor of a finding that

Defendants waived the privilege in this case.

Assuming that the Plaintiff's version of how Defendants conducted their privilege review is accurate,^{EN8} the Defendants obtained the results of the agreed-upon ESI search protocol and ran a keyword search on the text-searchable files using approximately seventy keywords selected by M. Pappas and two of his attorneys. Defendants, who bear the burden of proving that their conduct was reasonable for purposes of assessing whether they waived attorney-client privilege by producing the 165 documents to the Plaintiff, have failed to provide the court with information regarding: the keywords used; the rationale for their selection; the qualifications of M. Pappas and his attorneys to design an effective and reliable search and information retrieval method; whether the search was a simple keyword search, or a more sophisticated one, such as one employing Boolean proximity operators;^{EN9} or whether they analyzed the results of the search to assess its reliability, appropriateness for the task, and the quality of its implementation. While keyword searches have long been recognized as appropriate and helpful for ESI search and retrieval, there are well-known limitations and risks associated with them, and proper selection and implementation obviously involves technical, if not scientific knowledge. See, e.g., United States v. O'Keefe, 537 F.Supp.2d 14, 24 (D.D.C.2008) ("Whether search terms or 'keywords' will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics.... Given this complexity, for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread."); Equity Analytics, LLC v. Lundin, 248 F.R.D. 331, 333 (D.D.C.2008), ("[D]etermining whether a particular search methodology, such as keywords, will or will not be effective certainly requires knowledge beyond the ken of a lay person (and a lay lawyer)...."); ^{FN10}In re. Seroquel Prods. Liab. Litig., 244 F.R.D. 650, 660 n. 6, 662 (M.D.Fla.2007) (criticizing defendant's use of keyword search in selecting ESI for production, noting the failure of the defendant to provide information "as to how it organized its search for relevant material, [or] what steps it took to assure

reasonable completeness and quality control" and observing that "while key word searching is a recognized method to winnow relevant documents from large repositories ... [c]ommon sense dictates that sampling and other quality assurance techniques must be employed to meet requirements of completeness."); The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery, 8 Sedona Conf. J. 189, 194-95, 201-02 ("[A]lthough basic keyword searching techniques have been widely accepted both by courts and parties as sufficient to define the scope of their obligation to perform a search for responsive documents, the experience of many litigators is that simple keyword searching alone is inadequate in at least some discovery contexts. This is because simple keyword searches end up being both over- and under-inclusive in light of the inherent malleability and ambiguity of spoken and written English (as well as all other languages)."). Id. at 194-95. To address this known deficiency, the Sedona Conference suggests as best practice points, *inter alia*:

*6 Practice Point 3. The choice of a specific search and retrieval method will be highly dependent on the specific legal context in which it is to be employed.

Practice Point 4. Parties should perform due diligence in choosing a particular information retrieval product or service from a vendor.

Practice Point 5. The use of search and information retrieval tools does not guarantee that all responsive documents will be identified in large data collections, due to characteristics of human language. Moreover, differing search methods may produce differing results, subject to a measure of statistical variation inherent in the science of information retrieval.

Practice Point 6. Parties should make a good faith attempt to collaborate on the use of particular search and information retrieval methods, tools and protocols (including as to keywords, concepts, and other types of search parameters).

Practice Point 7. Parties should expect that their

choice of search methodology will need to be explained, either formally or informally, in subsequent legal contexts (including in depositions, evidentiary proceedings, and trials).

Id.; and [Information Inflation: Can the Legal System Adapt](#), 13 Rich. J.L. & Tech. 10, at *38, 40 (as cited at www.westlaw.com) (“[I]t is not surprising that lawyers and those to whom they delegate search tasks may not be particularly good at ferreting out responsive information through use of simple keyword search terms.... Accordingly, the assumption on the part of lawyers that any form of present-day search methodology will fully find ‘all’ or ‘nearly all’ available documents in a large, heterogeneous collection of data is wrong in the extreme.”).

Use of search and information retrieval methodology, for the purpose of identifying and withholding privileged or work-product protected information from production, requires the utmost care in selecting methodology that is appropriate for the task because the consequence of failing to do so, as in this case, may be the disclosure of privileged/protected information to an adverse party, resulting in a determination by the court that the privilege/protection has been waived. Selection of the appropriate search and information retrieval technique requires careful advance planning by persons qualified to design effective search methodology. The implementation of the methodology selected should be tested for quality assurance; and the party selecting the methodology must be prepared to explain the rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented. In this regard, compliance with the Sedona Conference Best Practices for use of search and information retrieval will go a long way towards convincing the court that the method chosen was reasonable and reliable, which, in jurisdictions that have adopted the intermediate test for assessing privilege waiver based on inadvertent production, may very well prevent a finding that the privilege or work-product protection was waived.

*7 In this case, the Defendants have failed to demonstrate that the keyword search they performed on the text-searchable ESI was reasonable.

Defendants neither identified the keywords selected nor the qualifications of the persons who selected them to design a proper search; they failed to demonstrate that there was quality-assurance testing; and when their production was challenged by the Plaintiff, they failed to carry their burden of explaining what they had done and why it was sufficient.

Further, the Defendants' attempt to justify what was done, by complaining that the volume of ESI needing review and time constraints presented them with no other choice is simply unpersuasive. Defendants were aware of the danger of inadvertent production of privileged/protected information and initially sought the protections of a non-waiver agreement such as that discussed in *Hopson, supra*. Had they not voluntarily abandoned their request for a court-approved non-waiver agreement, they would have been protected from waiver. Instead, they advised the court that they did not need this protection and elected to do a document-by-document privilege review. According to Defendants version of the facts, when they undertook an individualized review of the nontext-searchable ESI and determined that they could only review the title pages, they neither sought an extension of time from the court to complete an individualized review nor reinstated their request for a court-approved non-waiver agreement, despite their awareness of how it would have provided protection against waiver. In these circumstances, Defendants' protests that they did their best and that their conduct was reasonable rings particularly hollow.

The remaining factors to be assessed under the intermediate test may be quickly disposed of. The Defendants produced 165 asserted privileged/protected documents to the Plaintiff, so this case does not present an instance of a single document slipping through the cracks. Further, the court's *in camera* review of the documents reflects that many of them are email and other communications between the Defendants and their various attorneys, as well as draft discovery responses, documents relating to settlements in unrelated litigation, comments from M. Pappas to counsel regarding discovery responses, and email correspondence between M. Pappas and Ms. Turner, the ESI forensic expert retained by Defendants. Thus,

the disclosures were substantive-including numerous communications between defendants and their counsel. As noted by other district courts within the Fourth Circuit, any order issued now by the court to attempt to redress these disclosures would be the equivalent of closing the barn door after the animals have already run away. *FDIC v. Marine Midland Realty Credit Corporation*, 138 F.R.D. 479,483 (E.D.Va., 1991) (“Any order issued now by the court would have only limited effect; it could not force NBNE to forget what has already been learned”); *Parkway Gallery Furniture, Inc. v. Kittinger/Pennsylvania House Group*, 116 F.R.D. 46, 52 (M.D.N.C.1987) (“[W]hen disclosure is complete, a court order cannot restore confidentiality and, at best, can only attempt to restrain further erosion.”). And, while the precise dates of the disclosures of the documents at issue are not clear from the record-since the Defendants made a series of ESI productions over a several week period-it is noteworthy that the Defendants did not discover the disclosure, but rather the Plaintiff made the discovery and notified the Defendants that potentially privileged/protected ESI had been produced. Therefore, this is not an instance in which a party inadvertently produced privileged information to an adversary, discovered the disclosure promptly, and then took immediate steps to inform the adversary that they had received the information inadvertently, thus demanding that it be returned.

*8 While Defendants' counsel did assert privilege and inadvertent production promptly after being notified by the Plaintiff of the production of possible privileged/protected information, the more important period of delay in this case is the one-week period between production by the Defendants and the time of the discovery by the Plaintiff of the disclosures-a period during which the Defendants failed to discover the disclosure. Finally, the Defendants have pointed to no overriding interests in justice that would excuse them from the consequences of producing privileged/protected materials. The Plaintiff is blameless, but the Defendants are not, having failed to take reasonable precautions to prevent the disclosure of privileged information, including the voluntary abandonment of the non-waiver agreement that the Plaintiff was willing to sign. Every waiver of the attorney-client privilege produces unfortunate consequences for the party that disclosed the

information. If that alone were sufficient to constitute an injustice, there would never be a waiver. The only “injustice” in this matter is that done by Defendants to themselves. *Marine Midland Realty Credit Corp.*, 138 F.R.D. at 483 (“It is seldom ‘fundamentally unfair’ to allow the truth to be made public, and under the circumstances ... the Court finds that it would not be fair to reward Rowe's carelessness [in disclosing privileged materials] with a protective order.”). Accordingly, even under the intermediate test, the Defendants are not insulated from waiver.

Sufficiency of Defendants' Assertion of Privilege/Protection

In addition to arguing that the Defendants waived any privileged or protected status that the disclosed documents had, the Plaintiff also appears to contend that the Defendants failed to properly establish the existence of the privilege and protection asserted in the first instance. Specifically, the Plaintiff argues that despite the requirements of Fed.R.Civ.P. 26(b)(5) and Discovery Guideline 5.c of this court that claims of privilege must be particularized, the Defendants failed to meet this obligation, and further, failed to comply with an order of this court ^{FN11} regarding the proper way to assert privilege in responding to discovery requests. See Pl.'s Supp'l Mem. 1, 6, 10, Paper No. 221; Pl.'s Reply Mem. 1, Paper No. 230. While Defendants do not deny that they failed to comply with the court's order, they contend that when the parties met to confer regarding the sufficiency of the Defendants' privilege/protection claims, the parties “essentially” came to an agreement that for all but eleven of the 165 documents at issue, the asserted privilege/protection was legitimate.^{FN12} Accordingly, Defendants argue that if the waiver issue is decided in their favor, there are only eleven documents for which the assertion of privilege/protection is challenged by Plaintiff. Defs.' Opp'n Mem. 6-7, Paper No. 225. Having decided the waiver issue against the Defendants, there is no need to reach the question of whether privilege/protection was properly asserted in the first instance. However, given the recurring problems associated with resolving disputed privilege/protection claims during discovery, it would be helpful to state the procedures that need to be followed in this process for the benefit of future cases.

*9 While the scope of discovery in civil cases broadly encompasses facts relevant to the claims and defenses raised in the pleadings, and, on a showing of good cause, may even be extended to facts relevant to the subject matter of the litigation, [Fed.R.Civ.P. 26\(b\)\(1\)](#), it does not include privileged information. *Id.* Similarly, work-product protected information is beyond the reach of discovery unless the requesting party makes a showing of substantial need for the information and inability to obtain its substantial equivalent without undue hardship. [Fed.R.Civ.P. 26\(b\)\(3\)](#). Because the responding party is entitled to refuse to produce requested discovery if it is privileged or work product protected, the rules require that when doing so, the responding party must “describe the nature of the documents, communications, or tangible things not produced or disclosed-and do so in a manner that, without revealing information itself privileged or protected, will enable other parties to assess the claim.” [Fed.R.Civ.P. 26\(b\)\(5\)\(A\)\(ii\)](#). This requirement was added in the 1993 amendments to the rules of civil procedure, and in the words of the advisory committee:

The party [asserting privilege/protection] must also provide sufficient information to enable other parties to evaluate the applicability of the claimed privilege or protection. Although the person from whom the discovery is sought decides whether to claim a privilege or protection, the court ultimately decides whether, if this claim is challenged, the privilege or protection applies. Providing information pertinent to the applicability of the privilege or protection should reduce the need for in camera examination of the documents.

[Fed.R.Civ.P. 26](#) advisory committee's note. Neither the rule nor the advisory committee comment specifies exactly how the party asserting privilege/protection must particularize its claim. The most common way is by using a privilege log, which identifies each document withheld, information regarding the nature of the privilege/protection claimed, the name of the person making/receiving the communication, the date and place of the communication, and the document's general subject matter. *See, e.g.*, Discovery Guideline 9.c.; Paul W.

Grimm, Charles S. Fax, & Paul Mark Sandler, *Discovery Problems and Their Solutions*, 62-64 (2005)(“To properly demonstrate that a privilege exists, the privilege log should contain a brief description or summary of the contents of the document, the date the document was prepared, the person or persons who prepared the document, the person to whom the document was directed, or for whom the document was prepared, the purpose in preparing the document, the privilege or privileges asserted with respect to the document, and how each element of the privilege is met for that document.”).*Id.* at 62-63.

In actuality, lawyers infrequently provide all the basic information called for in a privilege log, and if they do, it is usually so cryptic that the log falls far short of its intended goal of providing sufficient information to the reviewing court to enable a determination to be made regarding the appropriateness of the privilege/protection asserted without resorting to extrinsic evidence or *in camera* review of the documents themselves. Few judges find that the privilege log is ever sufficient to make the discrete fact-findings needed to determine whether a privilege/protection was properly asserted and not waived. Further, because privilege review and preparation of privilege logs is increasingly handled by junior lawyers, or even paralegals, who may be inexperienced and overcautious, there is an almost irresistible tendency to be over-inclusive in asserting privilege/protection. While some of this tendency is understandable given the consequences of mistakenly producing privileged/protected information, the experience of many judges is that when the documents themselves are reviewed, it often turns out that a much smaller percentage of documents actually meet the requirements of the asserted privilege/protection than was claimed by the asserting party. Counsel should be wary of filing a response to a Rule 34 document production request that asserts privilege/protection as a basis for refusing to make requested production without having a factual basis to support each element of each privilege/protection claimed for each document withheld, because doing so is a sanctionable violation of [Fed.R.Civ.P. 26\(g\)](#).

*10 Requesting parties also know of the limited utility of privilege logs (for they likely have served

similar privilege logs in response to their adversary's discovery requests), and thus, when they receive the typical privilege log, they are wont to challenge its sufficiency, demanding more factual information to justify the privilege/protection claimed. This, in turn, is often met with a refusal from the producing party, and it does not take long before a motion is pending, and the court is called upon to rule on the appropriateness of the assertion of privilege/protection, often with the producing party's "magnanimous" offer to produce the documents withheld for *in camera* review. *In camera* review, however, can be an enormous burden to the court, about which the parties and their attorneys often seem to be blissfully unconcerned.

For example, in order for the court to determine whether the attorney-client privilege was properly asserted regarding a particular document, the court must make the following fact determinations:

(1) the asserted holder of the privilege is or sought to become a client; (2) the person to whom the communication was made (a) is a member of the bar of a court, or his subordinate and (b) in connection with this communication is acting as a lawyer; (3) the communication relates to a fact of which the attorney was informed (a) by his client (b) without the presence of strangers (c) for the purpose of securing primarily either (i) an opinion on law or (ii) legal services or (iii) assistance in some legal proceeding, and not (d) for the purpose of committing a crime or tort; and (4) the privilege has been (a) claimed and (b) not waived by the client.

In re Allen, 106 F.3d 582, 600 (4th. Cir.1997) (holding also that "a district court's holding that the attorney-client privilege does not protect communications rest[s] essentially on determinations of fact"). *Id.* at 601. Sometimes the document itself makes this clear, such as when the lawyer writes to the client to provide an opinion and the correspondence reflects that it is a confidential communication. Often times, however, it is impossible to determine if the privilege applies without extrinsic evidence, which must be provided by affidavit, deposition transcript, or other source. The time it takes the court to review this extrinsic

evidence on a document-by-document basis can be extensive, particularly given the tendency of lawyers to be over-inclusive in the assertion of privilege/protection in the first place. It should go without saying that the court should never be required to undertake *in camera* review unless the parties have first properly asserted privilege/protection, then provided sufficient factual information to justify the privilege/protection claimed for each document, and, finally, met and conferred in a good faith effort to resolve any disputes without court intervention. *United States v. Zolin*, 491 U.S. 554, 571-72, 109 S.Ct. 2619, 105 L.Ed.2d 469 (1989) ("[W]e cannot ignore the burdens *in camera* review places upon the district courts, which may well be required to evaluate large evidentiary records without open adversarial guidance by the parties.... Before engaging in *in camera* review ... the judge should require a showing of a factual basis adequate to support a good faith belief by a reasonable person'... that *in camera* review of the materials may reveal evidence to establish the claim [of privilege/protection]".) (internal citations omitted); *United States v. Family Practice Assocs. of San Diego*, 162 F.R.D. 624, 627 (S.D.Ca.1995) ("Prior to an *in camera* review there must first be a sufficient evidentiary showing of a legitimate issue as to application of a privilege or other protection. *In camera* review should not replace effective adversarial testing of the claimed privileges and protection."); *Diamond State Ins. Co. v. Rebel Oil Co.*, 157 F.R.D. 691, 700 (D.Nev.1994) ("In camera review is generally disfavored. It is not to be used as a substitute for a party's obligation to justify its withholding of documents. In camera review should not replace the effective adversarial testing of the claimed privileges and protections. Resort to *in camera* review is appropriate only after the burdened party has submitted detailed affidavits and other evidence to the extent possible.") (internal citations omitted); and *Caruso v. Coleman Co.*, 1995 WL 384602, at * 1 ("[R]esort to *in camera* review is appropriate only *after* the burdened party has submitted detailed affidavits and other evidence to the extent possible. Unfortunately, this court is put in the untenable position of having to speculate in order to determine which privileges apply to the individual documents. A court has the right to refuse to engage in such speculation, since the burden of proving the attorney-client or work-product privileges rests on the

party claiming the privilege.”) (internal citations omitted) (emphasis in original); Weber v. Paduano, No. 02 Civ. 3392, 2003 WL 161340, at *14 (S.D.N.Y. Jan. 22, 2003) (holding *in camera* review should not be undertaken routinely, but only after the party asserting privilege has submitted an adequate record to support the claim); Bowne of New York City v. AmBase Corp., 150 F.R.D. 465, 475 (S.D.N.Y.1993) (“AmBase’s suggestion of *in camera* review in lieu of an evidentiary presentation is misplaced. Such review ... is not, however, to be routinely undertaken, particularly in a case involving a substantial volume of documents, as a substitute for a party’s submission of an adequate record in support of its privilege claims.”); and Conde v. County of Suffolk, 121 F.R.D. 180, 190 (E.D.N.Y.1988) (“After giving defendants an opportunity to respond (with a brief and possible supplemental affidavits or declarations), the court can determine whether the defendants have made the requisite threshold showing to invoke the privilege. If the court finds that the defendant has not satisfied its threshold burdens, direct disclosure is in order. If the threshold burdens are met, the court may then review the materials at issue *in camera* and decide which, if any to withhold from disclosure.”).

*11 All of this has led to some fairly strongly worded statements from courts about what a party must do to substantiate its claim of privilege or protection. See Parkway Gallery Furniture, Inc. v. Kittinger/Pennsylvania House Group, 116 F.R.D. 46, 48 (M.D.N.C.1987) (“Disputes over whether the attorney-client privilege has been waived through inadvertent production of the documents or on the basis of the fraud or crime exception to the privilege often involve contested facts necessitating an evidentiary showing. Generally, the proponent or party claiming rights or benefit of an assertion bears the burden of establishing his contention.”); Caruso, 1995 WL 384602, at *1 (“A general allegation of privilege is insufficient. Instead, a clear showing must be made which sets forth the items or categories objected to and the reason for that objection. Accordingly, the proponent must provide the court with enough information to enable the court to determine the privilege, and the proponent must show by affidavit that precise facts exist to support the claim of privilege.”); United States v. Burns, 162 F.R.D. 624, 627-28 (S.D.Cal.1995) (finding

Defendant’s failure to make out a factual showing by “detailed affidavits or other evidence” waived privilege); Nutmeg Ins. Co. V. Atwell, Vogel & Sterling, 120 F.R.D. 504, 510 (W.D.La.1988) (“In considering whether a proponent of the privilege is entitled to protection, the courts must place the burden of proof squarely upon the party asserting privilege. Accordingly, the proponent must provide the court with enough information to enable the court to determine privilege, and the proponent must show by affidavit that precise fact exist to support the claim of privilege.”); Church of Scientology Intern. v. U.S. Dept. of Justice, 30 F.3d 224, 231 (1st Cir.1994) (“These declarations are written too generally to supplement the index in any meaningful way.... Thus, none of the functions of the index ... are served: the declarations do not demonstrate careful analysis of each document by the government; the court has not been assisted in its duty of ruling on the applicability of an exemption; and the adversary system has not been visibly strengthened.”); United States v. First State Bank, 691 F.2d 332, 335 (7th Cir.1982) (“A taxpayer need not reveal so many facts that the privilege becomes worthless but he must at least identify the general nature of that document, the specific privilege he is claiming for that document, and facts which establish all the elements of the privilege he is claiming. These allegations must be supported by affidavits.”); In re French, 162 B.R. 541, 548 (Bankr.D.S.D.1994) (“Debtor did, indeed, file an affidavit claiming the privilege, but the timing was off, and, even then, it was in the form of a ‘blanket’ assertion rather than articulated specific facts giving rise to a privilege.”).

Thus, insuring that a privilege or protection claim is properly asserted in the first instance and maintained thereafter involves a several step process. First, pursuant to Fed.R.Civ.P. 26(b)(5), the party asserting privilege/protection must do so with particularity for each document, or category of documents, for which privilege/protection is claimed. At this first stage, it is sufficient to meet the initial burden by a properly prepared privilege log. If, after this has been done, the requesting party challenges the sufficiency of the assertion of privilege/protection, the asserting party may no longer rest on the privilege log, but bears the burden of establishing an evidentiary basis-by affidavit, deposition transcript, or other evidence-for

each element of each privilege/protection claimed for each document or category of document. A failure to do so warrants a ruling that the documents must be produced because of the failure of the asserting party to meet its burden. If it makes this showing, and the requesting party still contests the assertion of privilege/protection, then the dispute is ready to submit to the court, which, after looking at the evidentiary support offered by the asserting party, can either rule on the merits of the claim or order that the disputed documents be produced for *in camera* inspection.

***12** In this case, the court made it clear to the Defendants that they were obligated to follow this procedure. *See* Letter Order, Dec. 28, 2007, Paper No. 194, and Pl.'s Letter of Feb 20, 2007 at 3, Paper No. 212, (citing to the Transcript of Dec. 21, 2007 Telephone Hearing, pp 16-17, when the court outlined the procedures to be used when submitting a privilege log). The Plaintiff argues that Defendants failed to comply with the court's order, and the Defendants have not demonstrated that they did. Had I not ruled that any privilege/protection already was waived, then the effect of a failure by the Defendants to comply with the court's order regarding the proper manner in which to assert privilege/protection would have warranted an order to produce the materials for failure to carry the burden of demonstrating the existence of the privilege/protection claimed.

Conclusion

For the reasons stated, the court finds that the Defendants waived any privilege or work-product protection for the 165 documents at issue by disclosing them to the Plaintiff. Accordingly, the Plaintiff may use these documents as evidence in this case, provided they are otherwise admissible. In this regard, the Plaintiff has only sought use of the documents themselves, and the court has not been asked to rule, and accordingly does not, that there has been any waiver beyond the documents themselves.

FN1. The 165 documents were produced to me for review *in camera*. Having done so, it is apparent that many do not qualify as attorney-client privileged or work-product protected. For example, the following

documents were asserted to be privileged or protected, yet the court's *in camera* review discloses that these assertions are without merit: Doc. No. 18 (discovery request from Plaintiff to Defendant); Doc. Nos. 28, 32 (email between employee of **CreativePipe** to M. Pappas, not discussing any materials that legitimately could be characterized as confidential); Doc.Nos. 24, 60 (email from Plaintiff's attorney to Defendants' attorney); Doc. Nos. 56, 61-65 (email between M. Pappas and G. Turner, Defendants' ESI expert, regarding payment); Doc. Nos. 105, 111, 130-133, 148-149, 151-158 (pictures of products, such as benches, trash can); Doc. No. 143 (page from invoice M. Pappas from attorney, no confidential information contained). It should be noted that the Defendants' failure to comply with the court's order of December 28, 2007, Paper No. 194, regarding how to handle assertion of privilege/protection claims resulted in an absence from the record of the factual basis to support their claims.

FN2. This court informed Defendants that they had the burden of providing an evidentiary basis to establish each element of the attorney-client privilege and work-product protection for each document at issue. Letter Order, Dec. 28, 2007, Paper No. 194. Notwithstanding, Defendants failed to do so, relying instead on the privilege logs that they provided to VSI, which did little more than briefly identify and describe each document and identify the basis for the refusal to produce it. As will be explained in this memorandum and order, when a party refuses to produce documents during discovery on the basis that they are privileged or protected, it has a duty to particularize that claim. [Fed.R.Civ.P. 26\(b\)\(5\), Discovery Guideline 9.c; Caruso v. Coleman Co., CIV. A. No. 93-CV-6733, 1995 WL 384602, at *1, \(E.D.Pa. June 22, 1995\); Bowne of New York City v. AmBase Corp., 150 F.R.D. 465, 474 \(S.D.N.Y.1993\); In re Pfohl Bros. Landfill Litig., 175 F.R.D. 13, 20 \(W.D.N.Y.1997\); United States v.](#)

Kovel, 296 F.2d 918, 923 (2d. Cir.1961). While a privilege log that complies with Discovery Guideline 9.c is an acceptable way to do so initially, once the claims of privilege/protection have been challenged by the requesting party, the producing party must then establish an evidentiary basis to support the privilege/protection claim. Failure to do so results in a forfeiture of the privilege/protection claimed. Bowne, 150 F.R.D. at 474 (holding that if the party claiming privilege fails to provide sufficient detail to demonstrate all legal requirements to make out the privilege, the claim must be rejected); Fox v. California Sierra Fin. Servs., 120 F.R.D. 520, 524 (N.D.Cal.1998) (finding that a party claiming privilege as basis for withholding discovery must properly identify each document and the basis for the privilege claimed); In re Pfohl Bros., 175 F.R.D. at 20 (holding “[m]ere conclusory or *ipse dixit* assertions of privilege” fail to satisfy the burden of demonstrating the applicability of a privilege).

FN3. In *Hopson*, this court discussed the dangers inherent in using non-waiver agreements, such as “clawback” or “quick-peek” agreements, and noted that reliance on them could nonetheless result in a determination that privilege and work-product protection had been waived, notwithstanding the agreement, given the current state of the substantive law regarding privilege waiver. Hopson, 232 F.R.D. at 236-38. The court further identified a process that could be employed within the boundaries of existing privilege waiver law that would significantly improve the likelihood of avoiding privilege waiver. The court noted:

[I]t is essential to the success of this approach in avoiding waiver that the production of inadvertently produced privileged electronic data must be at the compulsion of the court, rather than solely by the voluntary act of the producing

party, and that the procedures agreed to by the parties and ordered by the court demonstrate that reasonable measures were taken to protect against waiver of privilege and work product protection.

Id. at 240. Defendants' counsel were aware of the requirements of *Hopson*. Pl.'s Letter of Mar. 29, 2007, Paper No. 79. The court's request for additional briefing regarding the burdens associated with conducting privilege review within the time allotted for Defendants to produce the ESI to Plaintiff was aimed at developing a factual record that would permit a *Hopson* compliant non-waiver agreement to be approved by the court.

FN4. It also is worth noting that Defendants' current counsel, James Rothschild, of Anderson Coe and King LLP, and Joshua Kaufman, of Venable LLP entered their appearance after all the events that are relevant to resolving the pending dispute had taken place and are not responsible for any of the actions or inactions that contributed to the court's ruling.

FN5. None of the Fourth Circuit cases reviewed in *Hopson* examined privilege waiver in the context of a voluminous document production during discovery in a civil case, and none of them considered the extra challenges of preventing privilege waiver posed by handling voluminous production of ESI, which is a relatively new phenomenon. The advisory committee notes to recently amended Fed.R.Civ.P. 26(b)(5) acknowledge these challenges:

The Committee [on the Rules of Practice and Procedure] has repeatedly been advised that the risk of privilege waiver and the work necessary to avoid it, add to the costs and delay of discovery. When the review is of electronically stored information, the risk of waiver, and the time and effort required to avoid it, can increase substantially because of the

volume of electronically stored information and the difficulty in ensuring that all information to be produced has in fact been reviewed.

[Fed.R.Civ.P. 26](#) advisory committee's note. Notwithstanding this recognition, however, the recently adopted rules of civil procedure relating to ESI do not effect any change in the substantive law of privilege waiver, as was discussed in some detail in *Hopson*, *supra*, because the Rules Enabling Act precludes creation or abrogation of any privilege by ordinary rule making. This is reserved for Congress. [28 U.S.C. § 2074\(b\)](#) (1988). Following the *Hopson* decision, however, the Advisory Committee on the Rules of Evidence conducted hearings on this issue and, following public comment, proposed a new rule of evidence: Rule 502. The Committee approved the proposed rule and the Judicial Conference then forwarded it to Congress where it was passed by the Senate as S. 2450. It is still pending in the House of Representatives. If enacted by Congress, Proposed Federal Evidence Rule 502 would solve the problems *Hopson* discussed and protect against privilege waiver under circumstances similar to those presented in this case if the parties entered into a non-waiver agreement that meets the requirements of the proposed rule, and the court, in turn, approved it. Until this happens, however, the procedures identified in *Hopson* are the only ones that provide a possible means of avoiding waiver in those jurisdictions that have not recognized the intermediate approach to waiver by inadvertent production (and, as noted, the Defendants initially sought to enter a non-waiver agreement such as discussed in *Hopson*, but then abandoned this effort). Should the issue of privilege waiver by inadvertent production of voluminous ESI be considered by the Fourth Circuit at some time in the future, it may be hoped that the court will be

cognizant of the unique problems presented with regard to avoiding privilege waiver presented by ESI discovery, as well as the fact that the approval of Proposed Evidence Rule 502 by the Committee on the Rules of Evidence, as well as the Judicial Conference, recognizes a need to provide relief in this difficult area. The substantive law of privilege is not rigid and inflexible, [Hopson](#), 232 F.R.D. at 240 (citing [Jaffee v. Redmond](#), 518 U.S. 1, 8, 116 S.Ct. 1923, 135 L.Ed.2d 337 (1996)), but is governed by principles of the common law as interpreted "by the courts of the United States in the light of reason and experience." [Fed.R.Evid. 501](#). Experience has now shown that ESI discovery presents unique, heretofore unrecognized, risks of waiver of privilege or work-product protection even when the party asserting the privilege or protection has exercised care not to waive it. The approval of Proposed Evidence Rule 502 by the Judicial Conference is a reasoned response to this new experience, but still pending in Congress. For those courts that have yet to decide which approach to follow regarding the inadvertent disclosure of privileged material during ESI discovery, the commentary to the proposed rule is worthy of consideration.

FN6. As noted in [Continental Casualty Co. v. Under Armour, Inc.](#), 537 F.Supp.2d 761 (D.Md.2008), if documents qualify as both attorney-client privileged and work-product protected, separate analysis is required to determine whether inadvertent production constitutes waiver. However, the majority view is that disclosure of work-product material in a manner that creates a substantial risk that an adversary will receive it waives the protection. *Id.* at 772-73 (citing [Restatement \(Third\) of the Law Governing Lawyers § 91 \(2000\)](#)). In this case, Defendants' voluntary, though inadvertent, production of the 165 documents directly to counsel for the

Plaintiff waived any work-product protection they may have had. *Id.*

[FN7.](#) Citing *dicta* in [Hopson, 232 F.R.D. at 237 n. 27](#), Defendants argue that state privilege waiver law controls the determination of whether the inadvertent production of privileged ESI waived the privilege, at least as to the supplemental state law claims that have been pleaded by Plaintiff. Defs.' Opp'n Mem. 10 n. 4, Paper No. 225. And, as they correctly note, the Maryland Court of Special Appeals has adopted the intermediate test in [Elkton Care Center Associates, Ltd. Partnership v. Quality Care Management, 145 Md.App. 532, 805 A.2d 1177 \(2002\)](#). However, as this court more recently pointed out in [Continental Casualty Co., 537 F.Supp.2d at 768 n. 3](#), (citing cases), the majority of federal courts that have addressed the issue of what privilege law to apply in federal cases where both federal and state claims are pending, and where the law of privilege is different under federal law than it is under state law, have concluded that federal privilege law trumps state privilege law. If for no other reason than an appreciation of the shortness of life, a court ought not to be required to parse out competing outcomes under differing state and federal privilege law to apply to the same core facts presented in litigation that spawned both federal and state claims is a time consuming and challenging task. I agree that following the majority view is a better approach, and so adopt it in this decision. Consequently, federal privilege waiver law will apply to both the federal and state claims.

[FN8.](#) Which, on the record before me, the Defendants do not rebut.

[FN9.](#) Keyword searching may be accomplished in many ways. The simplest way is to use a series of individual keywords. Using more advanced search techniques, such as Boolean proximity operators, can enhance the effectiveness of

keyword searches. Boolean proximity operators are derived from logical principles, named for mathematician George Boole, and focus on the relationships of a "set" of objects or ideas. Thus, combining a keyword with Boolean operators such as "OR," "AND," "NOT," and using parentheses, proximity limitation instructions, phrase searching instructions, or truncation and stemming instructions to require a logical order to the execution of the search can enhance the accuracy and reliability of the search. *The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery*, 8 Sedona Conf. J. (2007) at 200, 202, 217-18 ("*Sedona Conference Best Practices*"); [Information Inflation: Can the Legal System Adapt?, 13 Rich. J.L. & Tech. 10 \(2007\)](#) at *37-41 (as cited at [www.westlaw.com](#)). In addition to keyword searches, other search and information retrieval methodologies include: probabilistic search models, including "Bayesian classifiers" (which searches by creating a formula based on values assigned to particular words based on their interrelationships, proximity, and frequency to establish a relevancy ranking that is applied to each document searched); "Fuzzy Search Models" (which attempt to refine a search beyond specific words, recognizing that words can have multiple forms. By identifying the "core" for a word the fuzzy search can retrieve documents containing all forms of the target word); "Clustering" searches (searches of documents by grouping them by similarity of content, for example, the presence of a series of same or similar words that are found in multiple documents); and "Concept and Categorization Tools" (search systems that rely on a thesaurus to capture documents which use alternative ways to express the same thought). See *Sedona Conference Best Practices, supra*, at 217-23.

[FN10.](#) The *O'Keefe* and *Equity Analytics* opinions have raised the eyebrows of some

commentators who have expressed the concern that they “engraft [\[Fed.R.Evid.\] 702](#) (and [\[Fed.R.Evid. 104\(a\)\]](#) into discovery ... [which, it is feared] would multiply the costs of discovery”, and, it is argued, this is a ‘path [that] is rife with unintended consequences”. See, e.g., [Rule 702 and Discovery of Electronically Stored Information](#), 8 Digital Discovery & E-Evidence (BNA) No. 5, at p. 3 (May 1, 2008). A careful reading of *O’Keefe* and *Equity Analytics*, however, should allay these concerns. In neither case did the court expressly hold that [Fed.R.Evid. 702](#) and [104\(a\)](#) were “engrafted” into the rules of discovery in civil proceedings (indeed, neither opinion even mentions [Rule 104\(a\)](#)). Instead, Judge Facciola made the entirely self-evident observation that challenges to the sufficiency of keyword search methodology unavoidably involve scientific, technical and scientific subjects, and *ipse dixit* pronouncements from lawyers unsupported by an affidavit or other showing that the search methodology was effective for its intended purpose are of little value to a trial judge who must decide a discovery motion aimed at either compelling a more comprehensive search or preventing one. Certainly those concerned about the *O’Keefe* and *Equity Analytics* opinions would not argue that trial judges are not required to make fact determinations during discovery practice. Indeed, such fact determinations inundate them. For example, deciding whether ESI discovery is not reasonably accessible because of undue burden or cost ([Fed.R.Civ.P. 26\(b\)\(2\)\(B\)](#)) involves factual determinations, as does determining whether discovery sought is too expensive or burdensome under [Fed.R.Civ.P. 26\(b\)\(2\)\(C\)](#); determining whether sanctions should be imposed for failing to preserve ESI or if the loss was a result of the routine, good faith operation of an electronic information system under [Fed.R.Civ.P. 37\(e\)](#); or determining whether documents withheld from disclosure are privileged or protected. Certainly the court is entitled to reliable factual information on which to

make such rulings. It cannot credibly be denied that resolving contested issues of whether a particular search and information retrieval method was appropriate in the context of a motion to compel or motion for protective order involves scientific, technical or specialized information. If so, then the trial judge must decide a method's appropriateness with the benefit of information from some reliable source—whether an affidavit from a qualified expert, a learned treatise, or, if appropriate, from information judicially noticed. To suggest otherwise is to condemn the trial court to making difficult decisions on inadequate information, which cannot be an outcome that anyone would advocate. For example, in the analogous technical area of sampling ESI, courts have recognized the need to have expert assistance to develop a valid random sampling protocol. See, e.g., [In re Vioxx Products Liability Litigation](#), No. 06-30378, 06-30379 2006 W.L. 1726675, at *2 n. 5 (5th Cir. May 26, 2006) (“By random sampling, we mean adhering to a statistically sound protocol for sampling documents.... The parties must provide expert assistance to the district court in constructing any protocol.”); [Manual for Complex Litigation \(Fourth\) § 11.446 \(2004\)](#) (“The complexity and rapidly changing character of technology for the management of computerized materials may make it appropriate for the judge to ... call on the parties to provide the judge with expert assistance, in the form of briefings on the relevant technological issues.”). Indeed, it is risky for a trial judge to attempt to resolve issues involving technical areas without the aid of expert assistance. In [American National Bank & Trust Co. v. Equitable Life Assurance Society](#), 406 F.3d 867, 879 (7th Cir.2005), the court reversed a magistrate judge's sanctions ruling that was predicated on sampling methodology the judge developed, and which the appellate court characterized as “arbitrary” and lacking “logical foundation.”

Moreover, if the court is to be given scientific or technical information to resolve a contested discovery matter, what standards should govern its evaluation? Should the court ignore a purported ESI expert's lack of qualifications if that shortcoming is demonstrated by the party opposing his opinion? Should the court accept opinions shown to be unsupported by sufficient facts or based on demonstrably unreliable methodology? The answer is obviously "No." Viewed in its proper context, all that *O'Keefe* and *Equity Analytics* required was that the parties be prepared to back up their positions with respect to a dispute involving the appropriateness of ESI search and information retrieval methodology-obviously an area of science or technology-with reliable information from someone with the qualifications to provide helpful opinions, not conclusory argument by counsel. The goal of [Federal Rule of Evidence 702](#) is to set standards to determine whether information is "helpful" to those who must make factual determinations involving disputed areas of science, technology or other specialized information. The rule is one of common sense, and reason-opinions regarding specialized, scientific or technical matters are not "helpful" unless someone with proper qualifications and adequate supporting facts provided such an opinion after following reliable methodology. That these common sense criteria are found in the rules of evidence does not render them off-limits for consideration during discovery. It is not unusual for pretrial factual determinations in civil cases to look to the Federal Rules of Evidence for assistance in resolving fact disputes. Indeed, in summary judgment practice, [Fed.R.Civ.P. 56\(e\)](#) requires that the parties support their motions with "such facts as would be admissible in evidence." 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. Should their selection be challenged by their adversary, and the court be called upon to make a ruling, then they should expect to support their position with affidavits or other equivalent information from persons with the requisite qualifications and experience, based on sufficient facts or data and using reliable principles or methodology.

For those understandably concerned about keeping discovery costs within reasonable bounds, it is worth repeating that the cost-benefit balancing factors of [Fed.R.Civ.P. 26\(b\)\(2\)\(C\)](#) apply to all aspects of discovery, and parties worried about the cost of employing properly designed search and information retrieval methods have an incentive to keep the costs of this phase of discovery as low as possible, including attempting to confer with their opposing party in an effort to identify a mutually agreeable search and retrieval method. This minimizes cost because if the method is approved, there will be no dispute resolving its sufficiency, and doing it right the first time is always cheaper than doing it over if ordered to do so by the court. Additionally, cost can be minimized by entering into a court-approved agreement that would comply with *Hopson*, or if enacted, Proposed Evidence Rule 502. In addition, there is room for optimism that as search and information retrieval methodologies are studied and tested, this will result in identifying those that are most effective and least expensive to employ for a variety of ESI discovery tasks. Such a study has been underway since 2006, when the National Institute of Standards and Technology (NIST), an agency within the U.S. Department of Commerce,

--- F.Supp.2d ----

--- F.Supp.2d ----, 2008 WL 2221841 (D.Md.)

(Cite as: --- F.Supp.2d ----, 2008 WL 2221841 (D.Md.))

embarked on a cooperative endeavor with the Department of Defense to evaluate the effectiveness of a variety of search methodologies. This project, known as the Text Retrieval Conference (TREC), evolved into the Trec LegalTrack, a research effort aimed at studying the e-discovery review process to evaluate the effectiveness of a wide array of search methodologies. This evaluative process is open to participation by academics, law firms, corporate counsel and companies providing ESI discovery services. See: <http://trec-legal.umiacs.umd.edu>. The next test will occur in the summer of 2008. The goal of the project is to create industry best practices for use in electronic discovery. This project can be expected to identify both cost effective and reliable search and information retrieval methodologies and best practice recommendations, which, if adhered to, certainly would support an argument that the party employing them performed a reasonable ESI search, whether for privilege review or other purposes.

FN11. See, e.g., Letter Order, Dec. 28, 2007, Paper No. 194 (“[W]ithin 30 days, Defendant[s] shall provide to Plaintiff an affidavit or other similar evidentiary support to establish each element of each privilege or work product protection asserted for each document for which privilege or work product is claimed.”).

FN12. As noted *supra*, footnote 1, the court's *in camera* review of the documents confirmed that there were numerous documents for which no legitimate claim of privilege or protection could be sustained.

D.Md.,2008.

Victor Stanley, Inc. v. Creative Pipe, Inc.

--- F.Supp.2d ----, 2008 WL 2221841 (D.Md.)

END OF DOCUMENT

Biographies

**Reza Alexander, Litigation & Practice Support Manager
DLA Piper UK LLP**



Reza Alexander is the Litigation & Practice Support Manager for DLA Piper UK, with over 13 years experience in leading and directing the delivery of cost effective technology solutions, document management and litigation support services in the legal industry.

An internationally recognised electronic evidence and case management expert, he is primarily responsible for advising on and developing best-practices in relation to cost effective electronic data collection, in house electronic data processing and industry leading techniques for cost and time effective review of voluminous electronic data both for disclosure and regulatory investigations to the firm's EMEA offices and clients.

Reza also serves as a member of DLA Piper's Electronic Discovery Readiness and Response Group, is an Editorial Board Member of the Litigation Support Today magazine, and a founding and active member of both the Litigation Support Technology Group [LiST] and the Sedona Conference WG6: International Electronic Information Management, Discovery and Disclosure Group.

Reza's forte is in sourcing and identifying cost-effective and practical solutions to the inherent challenges of volume litigation and regulatory investigations - leveraging technology to the clients' best advantages - and is particularly keen in the advancement and wider education of legal technology in the industry.

Reza is a frequent speaker at electronic disclosure and industry specific conferences and seminars nationally and internationally, and is a published author of various articles on electronic disclosure.

**Simon Attfield, Snr. Research Fellow
University College London**



Simon Attfield is a Senior Research Fellow at UCL Interaction Centre at University College London.. He has a background in Philosophy, Psychology, Information Science and Human Computer Interaction. His PhD thesis 'Information seeking, gathering and review: Journalism as a case study for the design of search and authoring systems' was awarded 'Highly Commended' in the The European Foundation for Management Development and Emerald Outstanding Doctoral Research Awards (2005)

(Information Science class).

Simon is currently working on the project 'Making Sense of Information' funded by the EPSRC. His research interests lie in the area of understanding information interaction in naturalistic settings and how the processes involved in sensemaking can be better supported. He has conducted numerous field studies of information behaviour, including studies in national news organizations (The Times, ITN), legal firms (Richards Butler, Freshfields Bruckhaus Deringer) and various healthcare settings. He has consulted to news, legal and medical information providers, published internationally in academic outlets, and presented research internationally to academic and commercial audiences.

**Jason R Baron, Director of Litigation
US National Archives and Records Administration**



Jason R. Baron has served since the year 2000 as Director of Litigation for the National Archives and Records Administration. Between 1988 and 1999, Mr. Baron served as trial attorney and senior counsel at the Department of Justice, defending the government's interests in complex federal court litigation, including in cases involving the preservation of White House email. He currently represents NARA on the Sedona Conference Working Group on Electronic Records Retention and Production, where he is member of the Steering Committee and Editor-in-Chief of the Sedona Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery¹. He also is a founding co-coordinator of the National Institute of Standards and Technology TREC legal track⁶, a multi-year international information retrieval project devoted to evaluating search issues in a legal context. Mr. Baron has been a Visiting Scholar at the University of British Columbia, and is currently an Adjunct Professor at the University of Maryland's College of Information Studies. He also presently serves on the Georgetown University Law Center Advanced E-Discovery Institute Advisory Board. His co-authored article, Information Inflation: Can the Legal System Adapt?, in the RICHMOND JOURNAL OF LAW AND TECHNOLOGY, is available online. Mr. Baron received his B.A. from Wesleyan University, and his J.D. from the Boston University School of Law.

**Robert S. Bauer, Chief Technology Officer
H5**



Bob Bauer has over 30 years of leadership in turning innovative technologies into strategic advantages. Dr. Bauer is the former vice president and founding CTO of Xerox Global Services. He joined Xerox's Palo Alto Research Center (PARC) in its inaugural year of 1970. On PARC senior staff, he led the System Sciences Laboratory and created PARC's Advanced Systems Development lab. These organizations delivered socio-technical innovations that leveraged deep understanding of how people can better manage increasing amounts of information in the workplace. He has incubated and helped create many companies that leverage innovative technologies that help businesses be more productive.

Dr. Bauer earned MS and Ph.D. degrees in electrical engineering from Stanford University. He is a Fellow of the American Physical Society and has been an advisor for the National Academy of Sciences, the National Science Foundation, UNESCO, and the federal Departments of Commerce, Defense, and Homeland Security. Bauer also serves on the Board of Advisors for a number of start-ups as well as RPI and Penn State business schools and GTI Group, a technology venture capital firm.

¹ Sedona Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, available at www.thesedonaconference.org.

**Frank Bennett, Jr., Associate Professor of Law
Nagoya University**



J.D. (UCLA, 1987). Lecturer in Law, SOAS (University of London), 1988 – 1998. Associate Professor of Law, Nagoya University, 1998 – present. Representative publications: Building Ownership in Modern Japanese Law: Origins of the Immobile Home, 26 Law in Japan (2000); Secondhand Japan: Used Goods Regulation 1645 – Present, 21&22 ZJapanR 37-53, 128-144 (2006); An “Open Source” Model for Statutory Translation (paper presented to the first Copymart conference, Meijo University, March 9, 1999).

**Ian Black, Head of Global Operations
Autonomy Group**



Ian Black joined Autonomy in June 2000 as Director of Corporate Communications and has served as Managing Director Aungate since 2003. Prior to joining Autonomy Ian operated a number of his own businesses providing communications consultancy to a distinguished list of multi-nationals and European government agencies. After joining British Aerospace - latterly BAE SYSTEMS - in 1989 and as Head of Corporate Communications, Ian achieved a string of world 'firsts' including deployment of the world's first live business television system which was itself a forerunner of another world first: a global automated Intranet. Throughout the 90's Ian's work earned over a dozen awards as part of BAE's top-120 group driving cultural change throughout the company. He also founded and participated in a number of external I.T networks and organizations such as the DTI's CSSA National E-Commerce Committee. Ian studied Physiology at Henley Management College and Business Administration at Filton College Bristol.

**Ann Blandford, Professor
University College London**



Ann Blandford is Professor of Human–Computer Interaction and Director of UCL Interaction Centre. Her research spans theories and methods for evaluating interactive systems. An important focus for her work has been on understanding the user experience with digital libraries, and how people work with and make sense of information. She has led several research projects in this area, covering questions from how people formulate queries to how they fit information finding into their broader information work. She has published over 200 papers on these are related themes.

**David T. Chaplin, Vice President of Advanced Search Technologies
Kroll Ontrack**



David Chaplin founded Engenium Corporation, a leading provider of conceptual search and clustering technologies. Engenium was recognized in 2006 and 2007 by KMWorld Magazine as one of the 100 Companies that Matter in Knowledge Management and has received Trend Setting Product awards over the same two year period, Dave served as President and CEO of Engenium until the company was acquired by Kroll Ontrack in November, 2006 where he is currently the Vice President of Advanced Search Technologies. Mr. Chaplin holds a B.A. in Economics from Hartwick College.

Chris Dale

The e-Disclosure Information Project



Chris Dale qualified as a solicitor in 1980 after reading History at Oxford.

Since 1993, he has been a consultant working with lawyers and with suppliers on e-Disclosure projects. His primary focus is on training and education aimed at raising awareness of the time and costs savings which e-disclosure brings and on the commercial and tactical advantages of being ready for litigation.

He runs the **e-Disclosure Information Project**, which aims to bring a consistent message to everyone – companies, lawyers, suppliers and judges – involved at any stage in the handling of discovery data. The Project is sponsored by companies, including LexisNexis, whose interest in electronic disclosure extends beyond merely selling products. In addition to speaking engagements, Chris maintains a web site and blog which are the most authoritative UK sources of objective information about e-disclosure.

Chris was previously a developer of litigation software and is an expert in conversion of data between litigation systems. Before that, he was a litigation partner in a large London firm of solicitors.

Carsten Görg, Postdoctoral Researcher

Georgia Institute of Technology



Carsten Görg is a postdoctoral researcher at the School of Interactive Computing at the Georgia Institute of Technology. He studied computer science and mathematics as a double major at Saarland University in Germany where he also received his Ph.D. degree in computer science. His main research is in the area of information visualization and visual analytics with a focus on building human-centered systems. He also applies visualization techniques to other domains and conducts research in the area of building recommendation systems for software developers.

Yunhyong Kim, Curation Resources Researcher

University of Glasgow



Yunhyong Kim is the Digital Curation Centre (DCC) Curation Resources Researcher at the Humanities Advanced technology and Information Institute (HATII), University of Glasgow, UK. Her research focuses on constructing experimental methodologies to test approaches to preservation planning and action with respect to digital collections, to ensure long term access, integrity and authenticity of digital objects within the collection. She holds a Phd in mathematics from the University of Cambridge and a master's degree in speech and language processing from the University of

Edinburgh. Her expertise is in automated processes related to language processing, information management, knowledge discovery, and system modelling. Before taking up her current post she specialised in automating the extraction of semantic metadata from digital material, as part of the ingest and appraisal processes related to digital repositories. She identified automated genre classification of digital documents as a key step in this process, and has isolated and tested key elements in measuring and comparing the robustness of genre classification systems. The results of her research have been presented at a range of conferences European Conference on advanced research in Digital Libraries (2006), the International

CODATA conference (2006), and the Hawaiian International Conference on System Sciences (2008), as well as the DELOS conferences on Digital Libraries (2007). She has published numerous papers resulting from her research some by Springer (Lecture Notes in Computer Science) and IEEE Computer Society Press.

Kelly KJ Kuchta, CEO
Forensics Consulting Solutions, LLC



KJ Kuchta founded Forensics Consulting Solutions in 2001. The firm has since become a leader in the electronic discovery field with recent projects including the Sprint/Nextel and the Proctor & Gamble/Gillette mergers. KJ has been a consultant for over 17 years. His background includes more than 10 years with Ernst & Young, First USA, and USAA Credit Card Center. In addition to conducting or managed over 5,000 investigations he is credited with introducing numerous groundbreaking ideas and concepts in the areas of risk management and investigations.

Mr. Kuchta received a Bachelor of Science in Criminal Justice from the University of Nebraska and an MBA from the University of Phoenix. He has testified as an expert in federal and state courts and is often called upon to speak on electronic discovery at conferences and association meetings nationwide.

Mounia Lalmas, Professor
Queen Mary, University of London



Mounia Lalmas is a Professor of Information Retrieval at Queen Mary, University of London, which she joined in 1999 as a lecturer. Prior to this, she was a Research Scientist at the University of Dortmund in 1998, a Lecturer from 1995 to 1997 and a Research Fellow from 1997 to 1998 at the University of Glasgow, where she received her PhD in 1996. Her research focuses on the development and evaluation of intelligent access to interactive heterogeneous and complex information repositories, and covering a wide range of domains such as HTML, XML, and MPEG-7. She co-led from 2002 to 2007 the international evaluation initiative for content-oriented XML retrieval (INEX), a large-scale project with over 80 participating organizations worldwide. She is currently the ACM SIGIR vice chair. She will take up a Microsoft Research/Royal Academy of Engineering Research Chair in Information Retrieval at the University of Glasgow in September 2008.

Stephen Mason, Associate Snr. Research Fellow
Institute of Advanced Legal Studies



Stephen Mason is barrister (www.stephenmason.eu) and a member of the IT Panel of the General Council of the Bar of England and Wales. He is the author and general editor of Electronic Evidence: Disclosure, Discovery & Admissibility (LexisNexis Butterworths, 2007) and International Electronic Evidence, (British Institute of International and Comparative Law, 2008). He is the author of Electronic Signatures in Law (Tottel, 2nd edn, 2007) and E-Mail, Networks and the Internet: A Concise Guide to Compliance with the Law (xpl publishing, 6th edn, 2006). He is the founder and general editor of the Digital Evidence and Electronic Signature Law Review.

**Chris May, CEO
IE Discovery, Inc.**



Chris May brings more than 15 years of experience in designing and developing Discovery Management technologies and litigation support systems to his role as CEO of IE Discovery.

While still pursuing undergraduate studies, Chris gained hands-on practical experience in litigation support working for the Austin-based firm of Graves, Dougherty, Hearon & Moody, P.C. Prior to earning his B.A. in Economics from the University of Texas in 1993, Chris signed on at IE Discovery—then a fledgling information technology consultancy with a two-person staff. Chris was appointed as President in 2001. Under his technological guidance, IE Discovery released InfoDox™ in 1999—the first Web-enabled discovery management solution designed specifically for the demand-driven litigation environment.

Chris has participated as a speaker at LegalTech and LawNet conferences, and he brings both hands-on practical knowledge and deep technical insight to IE Discovery's government and corporate lawyer clients. Chris has been published, and lectures on important legal technology topics such as optimizing electronic discovery, and designing corporate document and data retention policies.

**Douglas W. Oard, Associate Professor
University of Maryland**



Douglas W. Oard is Associate Dean for Research at the University of Maryland's College of Information Studies. He holds joint appointments as an associate professor in the College of Information Studies and the Institute for Advanced Computing Studies. Dr. Oard earned his Ph.D. in 1996 in Electrical Engineering from the University of Maryland, College Park in 1996, a Master of Electrical Engineering degree from Rice University in 1979, and a B.A. in Electrical Engineering and Mathematical Sciences in 1979. His research is focused on the design and evaluation of interactive systems to support information retrieval and sense-making in large collections of character-coded, scanned, and spoken language. He is best known for his work on cross-language information retrieval. Since 2006 he has helped to coordinate evaluation of information retrieval techniques for e-discovery in the Text Retrieval Conference's Legal Track².

**Jacki O'Neill, Research Scientist
Xerox Research Centre Europe**

Jacki O'Neill has worked as an ethnographer in the Work Practice Technology Group at XRCE since 2001. Her central area of interest lies in the design of useful, usable and innovative computer systems, through both the detailed understanding of work practices and a consideration of the interaction of the social and the technical in prototyping and development work. Her current research includes e-discovery, production printing and troubleshooting work. Though very different working environments there is a central theme of mediation: between human understandings and technology systems.

² National Institute of Standards and Technology TREC legal track
(see <http://trec-legal.umiacs.umd.edu>)

Ian Ruthven, Reader
University of Strathclyde



Ian Ruthven is a Reader in Information Seeking and Retrieval in the Department of Computer and Information Sciences at the University of Strathclyde. He currently heads the *i-lab* research group, an interdisciplinary research group centered on information and information technology with a broad portfolio of research including statistical data modeling, information retrieval, digital libraries, mobile information access, information strategy and public libraries.

His research is in the broad area of interactive information access; understanding how (and why) people search for information and how electronic systems might help them search more successfully. This includes modelling of interactive retrieval systems, user and technical evaluations, interface development and studies of systems in use. His most recent research has examined areas such as Personal Information Management, Complex Interactive Question Answering and the evaluation of novel information surrogates. His research has been funded by a variety of sources including EPSRC, the British Library, and the Royal Society of London.

Mark Sanderson, Reader
University of Sheffield



Mark Sanderson is a Reader at the University of Sheffield in the Information Studies Department. He is a researcher in information retrieval and is particularly interested in evaluation of search engines, but also work in geographic search, cross language IR, summarisation, image retrieval by captions, word sense ambiguity. He is the co-founder of the imageCLEF evaluation exercise and will be one of the PC Chairs for SIGIR 2009.

Jeane A Thomas, Partner
Crowell & Moring



Jeane A. Thomas is a partner in Crowell & Moring's Antitrust Group and the Co-Chair of the firm's E-Discovery practice.

Ms. Thomas is engaged in all types of antitrust representations, including mergers and joint ventures, class and individual civil litigation, and civil and criminal government investigations. She also counsels clients on a broad range of antitrust issues, including intellectual property and licensing issues, trade association law, the Hart Scott Rodino Act, and pricing and distribution issues. She has focused extensively on the telecommunications, technology, chemicals and healthcare/ pharmaceuticals industries.

In her role with the E-Discovery practice, Ms. Thomas has managed many types of E-Discovery matters in both government investigations and private litigation. She regularly counsels clients on Litigation Readiness Planning, including the development and application of effective document/data retention policies and corporate content policies, as well as E-Discovery response plans. Ms. Thomas is a member of the Sedona Conference Working Group on E-Discovery, and a member of the Advisory Board and Faculty of the Georgetown University Law Center Advanced Institute for E-Discovery. She regularly speaks and writes on E-Discovery issues.