Fall 12-15-2012

Application of Digital Forensic Science to Electronic Discovery in Civil Litigation

Brian Roux
bcroux@uno.edu

Application of Digital Forensic Science to Electronic Discovery in Civil Litigation

A Dissertation

Submitted to the Graduate Faculty of the

University of New Orleans

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in

Engineering and Applied Sciences

with an Emphasis in

Computer Science

and a Concentration in

Information Assurance

by

Brian Roux

B.S. University of New Orleans, 2007

M.S. University of New Orleans, 2008

December, 2012

Copyright

Dedication

I would like to dedicate this Doctoral dissertation to my parents and grandparents. Their continued support and encouragement made this long journey possible

Acknowledgement

Many people have been with me through this journey. While I cannot acknowledge everyone, I owe deepest gratitude to the following.

My dissertation committee members:

- Dr. Golden Richard, III – My major professor, friend, and fellow fountain pen aficionado.
- Dr. Vassil Roussev
- Dr. Jamie Nino
- Dr. Juliette Ioup
- Prof. Linda Sins

My parents, who have always been proud of me even when they had no idea what I was talking about.

To all my family and friends, for your support and encouragement.

Foreword

In some parts herein, my style of argument and discussion may seem overly formal or archaic. This is intentional. Because I discuss legal issues, situations, and cases in constructing my arguments I must address the legal paradox of "persuasive" arguments. In legal discourse, persuasive arguments rule the day - but persuasive arguments are not necessarily correct arguments or valid arguments. Many wrong things are persuasive for a time or within a certain subset of the populous. Therefore, when possible, I identify problems in legal contexts I am critiquing by identifying explicitly the logical fallacy committed.

Contents

List of Figures

List of Tables

Abstract

Following changes to the Federal Rules of Civil Procedure in 2006 dealing with the role of Electronically Stored Information, digital forensics is becoming necessary to the discovery process in civil litigation. The development of case law interpreting the rule changes since their enactment defines how digital forensics can be applied to the discovery process, the scope of discovery, and the duties imposed on parties. Herein, pertinent cases are examined to determine what trends exist and how they effect the field. These observations buttress case studies involving discovery failures in large corporate contexts along with insights on the technical reasons those discovery failures occurred and continue to occur.

The state of the art in the legal industry for handling Electronically Stored Information is slow, inefficient, and extremely expensive. These failings exacerbate discovery failures by making the discovery process more burdensome than necessary. In addressing this problem, weaknesses of existing approaches are identified, and new tools are presented which cure these defects. By drawing on open source libraries, components, and other support the presented tools exceed the performance of existing solutions by between one and two orders of magnitude. The transparent standards embodied in the open source movement allow for clearer defensibility of discovery practice sufficiency whereas existing approaches entail difficult to verify closed source solutions.

Legacy industry practices in numbering documents based on Bates numbers inhibit efficient parallel and distributed processing of electronic data into paginated forms. The failures inherent in legacy numbering systems are identified, and a new system is provided which eliminates these inhibiters while simultaneously better modeling the nature of electronic data which does not lend itself to pagination; such non-paginated data includes databases and other file types which are machine readable, but not human readable in format.

In toto, this dissertation provides a broad treatment of digital forensics applied to electronic discovery, an analysis of current failures in the industry, and a suite of tools which address the weaknesses, problems, and failures identified.

Keywords: Digital Forensics, Forensic Computer Science, Electronic Discovery, Electronically Stored Information, Discere, Black Friar

Introduction

Forensic science is the use of science to answer questions relevant to the law. Digital forensics encompasses the application of computer science to answer pertinent legal questions about computers, digital devices, data storage, and other similar systems. Mainstream perception of digital forensics typically encompasses crime and criminal cases, but the field is equally applicable to civil cases. Where criminal cases punish crime, civil cases compensate injuries.

The difference between civil and criminal cases is important because each poses different questions for forensic science to answer. The differences are more pronounced with digital forensics because where in criminal cases the questions posed will be answered at trial, in civil cases the legal interest will instead arise during discovery. This difference applies different rules to how digital forensics can operate, what data each side can gain access to, and ultimately what role practitioners will take on in a case.

Electronic data presents a problem for discovery. The computer revolution changed the way information is stored. Paper is increasingly supplanted by electronic documents, databases, and other information storage systems. The law was slow to recognize this shift, and only began to focus on the problem in 2006. The solutions put forth then were a response to problems cropping up in discovery. Despite being addressed by rule changes to the discovery process, the rules are like seeds which grow overtime as case law interprets them in the context of real cases. The way these rules are interpreted directly impacts digital forensics as a field, so it is vital to understand how the development came about as well as how it applies to cases a practitioner is involved in.

In this dissertation, I will identify the problems Electronically Stored Information ("ESI") presents in the context of discovery in civil litigation, identify the existing state of the art at use in industry for handling ESI, and I will demonstrate superior approaches to the problems presented. The explicit codification of rules dealing with ESI in the Federal Rules of Civil Procedure is a recent event, as of this writing, merely six years ago. The problems I seek to address here, therefore, first requires significant treatment to define this new area in digital forensic science in order to understand the problems and the solutions presented in context. To this end, the argument is organized into three parts; each part has a specific purpose behind it.

One of the main difficulties in applying scientific principles to the problems presented by ESI is that the

Law itself limits what solutions will be allowed. Courts, in considering whether a proposed measure should be allowed, examine a number of factors to determine whether a solution is reasonable especially in context of the specific facts, burdens, and costs. Because the analysis is fact specific, there is no quantified metric or formula to determine, a priori, whether a proposed solution is acceptable. Part I examines the origins of ESI in discovery, the FRCP changes in 2006 to codify ESI as an explicit part of discovery, and the subsequent cases which apply the revised FRCP to fact patterns in deciding whether discovery obligations have been met. By examining the development of the problems and the resulting new applications of forensic computer science, I define the parameters of this new field and argue the post-2006 cases show a continuing uncertainty regarding the standards courts will apply based on the underlying technologies containing ESI.

Discovery, as pre-trial procedure, is primarily driven by the parties themselves with the Courts involvement limited to resolving problems the parties cannot resolve amongst themselves. This paradigm diverges from other forensic contexts in that each side has an affirmative duty to produce the required information, but they are also assumed to be fulfilling that duty out of professional obligation. As such, the methodologies, protocols, and other information pertinent to demonstrating whether discovery obligations have been met are not accessible to an expert working for the opposing side. Instead, the opposing expert, in cases where the sufficiency or completeness of the discovery effort is being challenged, must rely on what information has been disclosed along with technical and scientific knowledge to detect failures. One might term this a meta-discovery analysis. Part II contains case studies for cases I was engaged as an expert in. The case studies involve three large corporations in the same industry with nearly identical failures in their discovery effort. Through the case studies, I demonstrate the difficulties inherent in challenging the sufficiency of a discovery effort, the problems encountered in meta-discovery, and the considerations identified in Part I shown in context. In my treatment of each case, I identify areas of ambiguity which are relevant for the various inquiries into reasonableness, burden, and cost the courts use in deciding the appropriateness of suggested solutions. By formalizing these ambiguous factors into a mathematical notation or model, I offer a way to apply scientific principles to the courts analysis for more rigorous weighting of a solution in terms of burden and cost.

Part III continues the formalization of ambiguous factors impacting burden and cost, and continues on to formalize the burden and cost of processing ESI into paginated formats in the manner currently used industry wide. With equations representing the burdens and costs of acquiring the ESI as well as processing the ESI for production to other parties, it is possible to compare new solutions in the same terms. In essence, with a formalized model for representing the factors a court must consider in evaluating discovery efforts, new solutions can be demonstrated rather than argued to be less burdensome and costly; the ability to demonstrate less burdensome and costly methods changes the legal analysis itself.

The current state of the art in the industry is examined, and problems in the primarily used tool are identified. New tools are presented which extend the state of the art by addressing the technical problems identified and which outperform current solutions by in excess of an order of magnitude. Further problems are identified in current practices for numbering documents which currently impede parallel and distributed processing approaches, and solutions are proposed which eliminate these impedances.

In sum, I examine the evolution of electronic discovery, provide case studies of matters I have been involved in which illustrate the abstract issues in a real context, and I introduce new tools which increase the efficiency of electronic discovery practices. Each of these topics is covered in sequence in an attempt to lay out both a broad analysis of the field as a whole in civil litigation, to provide insight into discovery disputes at a technical level often omitted from court decisions, to identify why these practices are so costly, and to demonstrate how better implementations can reduce cost and decrease turnaround time for data production in discovery.

Part I

Electronic Discovery

Chapter 1

The Zubulake Cases

Digital forensics' traditional magisterium is the detection, reconstruction, and identification of computer based crime with the intended purpose of its fruits being used to prosecute those responsible. Practitioners will continue to focus on the criminal arena, but the field as a whole must shift with equal fervor to electronic discovery in civil cases. This imperative stems from the amendments to the Federal Rules of Civil Procedure in 2006 which clarified the inclusion of Electronically Stored Information ("ESI") in the discovery process. To understand the differences facing the field in the civil arena, we must understand (1) how the discovery process has changed, (2) the difference in how and when digital forensic experts are employed, and (3) trends in how civil electronic discovery will be handled in the future.

The classic courtroom scene in any drama includes shocked gasps when a new piece of evidence, fact, document, or witness is produced at a pivotal moment thus altering the previously thought predestined outcome. These courtroom theatrics are why the modern discovery system was created – to allow for an ordered and just outcome to a case based on its merits.

> Contemporary civil discovery permits parties to compel the disclosure of witnesses, evidence, documents, and other matters before trial. Previous practice made available only a limited range of discovery devices only against some persons, only in some kinds of actions. Discovery was not always a matter of right, and was often limited to a scope narrower than the issues relevant at trial. This recipe could produce good courtroom drama, but not, critics argued, truth or justice.
>
> Modern discovery has changed that picture. Both state courts and the federal system have adopted broad civil discovery rules that permit a lawyer to uncover, in advance of trial, enormous amounts of information.[1]

Discovery is governed by the Federal Rules of Civil Procedure which provide the mechanism and framework for conducting an exchange of information relevant to the case between the involved parties.

[1] [50, at 415]

> The pre-trial deposition-discovery mechanism established by Rules 26 to 37 is one of the most significant innovations of the Federal Rules of Civil Procedure. Under the prior federal practice, the pre-trial functions of notice-giving issue-formulation and fact-revelation were performed primarily and inadequately by the pleadings. Inquiry into the issues and the facts before trial was narrowly confined and was often cumbersome in method. The new rules, however, restrict the pleadings to the task of general notice-giving and invest the deposition-discovery process with a vital role in the preparation for trial. The various instruments of discovery now serve (1) as a device, along with the pre-trial hearing under Rule 16, to narrow and clarify the basic issues between the parties, and (2) as a device for ascertaining the facts, or information as to the existence or whereabouts of facts, relative to those issues. Thus civil trials in the federal courts no longer need be carried on in the dark. The way is now clear, consistent with recognized privileges, for the parties to obtain the fullest possible knowledge of the issues and facts before trial.[2]

Hickman v Taylor[18] is a famous case which tested the limits of discovery with regard to attorney "work product" and privilege. Put shortly, there was an accident between a tug boat and a car float in which five of the nine crew members were drowned[3]. The four survivors gave testimony at a public hearing, but afterwards the attorney for the tug boat owners interviewed the survivors privately[4]. The question before the court was whether the plaintiff side could acquire this information through discovery, and in deciding it the court recognized a qualified privilege for an attorney's "work product"[5].

The establishment of the attorney work product doctrine occurred in 1947, the same year the ENIAC was put into service. The court's language revolved around files and memorandum; at that time there was no concept of what was coming in the revolution of the computer age. The work product doctrine was eventually included in the FRCP as rule 26(b)(3) which we will later see being applied in novel ways when determining what ESI is considered work product.[6].

There are differences between civil and criminal cases which impact the nature of an investigator's role and the difficulties inherent in the investigation. In a criminal case, evidence is directly examined then the methods and findings are presented at trial; it is common for both sides to examine the evidence directly especially if there are differences in expert opinions on what the evidence means. While similar at trial presentations occur in civil cases, this is separate from the discovery process which occurs in the pre-trial phase. Additionally, unlike where evidence is seized, in discovery the opposing side generally will not be given direct access to the data sources. Instead, data is 'produced' pursuant to agreements between the parties

[2] [50, at 415-16]
[3] [18, at 498]
[4] [18, at 507-09]
[5] [18, at 511-14]
[6] [32] (considering if database tags constitute work product.)

or orders of the court; it is an ethical obligation and affirmative duty for the attorneys to comply with the FRCP which governs discovery. As we will see, this lack of opportunity for mutual direct observation and verification that discovery is being conducted properly often leads to disputes. When disputes arise, it is difficult to prove a discovery abuse because the side which would raise the dispute does not have direct access to the total corpus of data to demonstrate the abuse; instead, they must use circumstantial evidence to demonstrate the reasonableness of their concerns.

Modern discovery encompasses ESI as a standard part of its process, but this was not always the case and the manner and magnitude of its incorporation has changed dramatically. There are three phases of significant change with regard to the discovery of ESI. The first phase came from the Zubulake decisions which provide an example of how detrimental badly handled electronic discovery can be to justice. The Zubulake decisions show how long and drawn out electronic discovery can be, how expensive it is, and how an entire case can hinge on a single email. The second phase came as a result of the Zubulake decisions; the FRCP were amended in 2006 to provide for and clarify a party's responsibilities regarding ESI from the anticipation of litigation to the advent of discovery. The 2006 rule changes had the effect of throwing the legal community into frenzy as it struggled to understand its obligations, as vendors popped up like mushrooms and snake oil salesmen to cure every woe, and the courts began hammering out precedent based on the new rules. The third phase is the phase we are presently in, where the legal community has more of a handle on electronic discovery, but where the courts have begun to be less forgiving of inadequate electronic discovery, and of overly adversarial practices.

I examine these three phases based on the circumstances which brought them about. First, I will examine the Zulbulake decisions (I, III, IV, and V) to show the issues the court was dealing with and how the courts understanding changed over the course of the decisions as the bad discovery practices came to light. Second, I will examine the 2006 amendments to the FRCP, which rules were changed and what those changes mean in the context of how digital forensic practitioners must precede. Finally, I will examine extensive case law which expands our understanding of how the courts apply the FRCP as amended in discovery.

Zubulake

1.1 Zubulake I

UBS Warburg hired Laura Zubulake ("Laura") in August 1999. Laura was hired to work on UBS' U.S. Asian Equities Sales Desk ("Desk") as a director and senior salesperson. Laura reported to Dominic Vail, who managed the Desk, and was told, at the time she was hired, that she would be considered to replace

Vail if he left. Vail left his position to join UBS' London office, but Laura was not considered to replace him. Instead, UBS hired Matthew Chapin to replace Vail. Laura, the only woman working on the Desk, alleged Chapin treated her differently than other Desk members. She alleged

> Chapin "undermined Ms. Zubulake's ability to perform her job by, inter alia: (a) ridiculing and belittling her in front of co-workers; (b) excluding her from work-related outings with male co-workers and clients; (c) making sexist remarks in her presence; and (d) isolating her from the other senior salespersons on the Desk by seating her apart from them." No such actions were taken against any of Zubulake's male co-workers.[7]

Laura filed a charge of gender discrimination with the Equal Employment Opportunity Commission on August 16, 2001 and was terminated on October 9, 2001. She filed suit pleading a number of claims. UBS filed a timely answer denying Laura's allegations.

Discovery commenced around June 3, 2002. Laura's first discovery request included request twenty-eight, "'[a]ll documents concerning any communication by or between UBS employees concerning Plaintiff.' The term document in Zubulake's request 'includ[es], without limitation, electronic or computerized data compilations.'" UBS produced approximately 350 pages of documents in response to Laura's request, including approximately 100 pages of emails. The production resulted in heated exchanges between the two parties; the parties later hammered out an agreement in conference with United States Magistrate Judge Gabriel W. Gorenstein. Party of the agreement touched on document request twenty-eight providing, "Defendants will [ ] ask UBS about how to retrieve e-mails that are saved in the firm's computer system and will produce responsive e-mails if retrieval is possible and Plaintiff names a few individuals."

> Pursuant to the 9/12/02 Agreement, UBS agreed unconditionally to produce responsive e-mails from the accounts of five individuals named by Zubulake: Matthew Chapin, Rose Tong (a human relations representation who was assigned to handle issues concerning Zubulake), Vinay Datta (a co-worker on the Desk), Andrew Clarke (another co-worker on the Desk), and Jeremy Hardisty (Chapin's supervisor and the individual to whom Zubulake originally complained about Chapin). UBS was to produce such e-mails sent between August 1999 (when Zubulake was hired) and December 2001 (one month after her termination), to the extent possible.[8]

UBS did not produce further emails, nor did it search its backup tape archives instead notifying Laura the search of backup tapes would be cost prohibitive. The first concrete inkling of wrongness occurred at this juncture when Laura noted in her objection to UBS' non-production, that she had produced over 450 pages

[7] [38, at 312]
[8] [38, at 313]

of correspondence to them, meaning a substantial portion of those conversations on the UBS side had been deleted or were otherwise missing.

1.1.1 UBS Backup Protocols & Email Systems

Judge Gorenstein ordered UBS to produce a witness to explain how UBS' email system, backup protocols, and retention policies worked. UBS produced its Manager of Global Messaging, Christopher Behny, to be disposed. Behny testified UBS had developed an extensive email backup system and preservation protocols, in part to comply with regulations imposed by the Securities and Exchange Commission. UBS used a redundant backup protocol for its email system: Email was archived to optical disk as well as to backup tape.

The tape based backup used Veritas NetBackup implementing a standard Grandfather-Father-Son rotation, using both incremental and full backups: "(1) daily, at the end of each day, (2) weekly, on Friday nights, and (3) monthly, on the last business day of the month. Nightly backup tapes were kept for twenty working days, weekly tapes for one year, and monthly tapes for three years. After the relevant time period elapsed, the tapes were recycled." Behny indicated the Veritas system used a snapshot based backup, rather than backing up email at the individual mailbox level; recovering an email required restoring the entire system from the snapshot. This would, in modern contexts, be considered a disaster recovery type backup system. Behny indicated each backup snapshot required five hours to restore.

The second backup system, the optical disk system, was applied to "registered traders" at UBS including those at the Desk. The optical system only recorded external email (going to or from an outside source), but did so immediately upon sending or reception, but not internal emails. The optical media was non-erasable, and non-rewritable, and UBS had an archive going back to 1998 when the system was launched. Emails recorded in the optical disk system were plain text searchable across body, headers, and other criteria.

1.1.2 The Court Considers Whether UBS Should Produce E-mails from Back-ups

In considering the whether to order discovery, the court relies on two observations (1) because of the way UBS' backup system is designed, UBS could search all its archived emails without restoring all the backup tapes, and (2) there are more pages of responsive emails produced by Laura (450) than the totality of UBS (100) leading the court to conclude there were more emails to discover. The court then proceeds to conduct a cost analysis while contemplating whether Laura and UBS should share the costs of the backup restoration. It is the court's analysis of cost analysis and accessibility which is important looking forward to the 2006

rule changes examined in the next chapter.

Whether electronic data is accessible or inaccessible turns largely on the media on which it is stored. Five categories of data, listed in order from most accessible to least accessible, are described in the literature on electronic data storage:

1. *Active, online data:* "On-line storage is generally provided by magnetic disk. It is used in the very active stages of an electronic records [sic] life-when it is being created or received and processed, as well as when the access frequency is high and the required speed of access is very fast, i.e., milliseconds." Examples of online data include hard drives.

2. *Near-line data:* "This typically consists of a robotic storage device (robotic library) that houses removable media, uses robotic arms to access the media, and uses multiple read/write devices to store and retrieve records. Access speeds can range from as low as milliseconds if the media is already in a read device, up to 10-30 seconds for optical disk technology, and between 20-120 seconds for sequentially searched media, such as magnetic tape." Examples include optical disks.

3. *Offline storage/archives:* "This is removable optical disk or magnetic tape media, which can be labeled and stored in a shelf or rack. Off-line storage of electronic records is traditionally used for making disaster copies of records and also for records considered 'archival' in that their likelihood of retrieval is minimal. Accessibility to off-line media involves manual intervention and is much slower than on-line or near-line storage. Access speed may be minutes, hours, or even days, depending on the access-effectiveness of the storage facility." The principled difference between nearline data and offline data is that offline data lacks "the coordinated control of an intelligent disk subsystem," and is, in the lingo, JBOD ("Just a Bunch Of Disks").

4. *Backup tapes:* "A device, like a tape recorder, that reads data from and writes it onto a tape. Tape drives have data capacities of anywhere from a few hundred kilobytes to several gigabytes. Their transfer speeds also vary considerably ... The disadvantage of tape drives is that they are sequential-access devices, which means that to read any particular block of data, you need to read all the preceding blocks." As a result, "[t]he data on a backup tape are not organized for retrieval of individual documents or files [because] ... the organization of the data mirrors the computer's structure, not the human records management structure." Backup tapes also typically employ some sort of data compression, permitting more data to be stored on each tape, but also making restoration more time-consuming and expensive, especially given the lack of uniform standard governing data compression.

5. *Erased, fragmented or damaged data:* "When a file is first created and saved, it is laid down on the [storage media] in contiguous clusters ... As files are erased, their clusters are made available again as free space. Eventually, some newly created files become larger than the remaining contiguous free space. These files are then broken up and randomly placed throughout the disk." Such broken-up files are said to be "fragmented," and along with damaged and erased data can only be accessed after significant processing.

The court pieces together its data accessibility analysis from a number of white papers, websites such as webopedia, unpublished manuscripts, and so forth. It identifies the first three types – (1) active, online data; (2) near-line data; and (3) offline/storage archive – as accessible media, and the last two types – (4) Backup tapes; and (5) Erased, fragmented or damaged data – as inaccessible. There are a number of errors and inconsistencies in the court's analysis.

First, the court refers to robotic storage devices such as tape libraries as belonging to its defined type 2, while referring to optical disks or magnetic tapes which can be stored on a shelf as type 3, and backup tapes as type 4. Each of the three categories mentioned references magnetic tape, but two of the categories are accessible and the last is inaccessible by the court's analysis. The court notes of backup tapes that they are not intended for restoring single files because, "data mirrors the computer's structure, not the human records management structure."[9] The court's articulated understanding of backup tape media is clearly defective. It has coupled the general physical media's properties with the backup schema's, and treated them as one immutable being.

The worst case for a backup tape would be the old *nix way[10] of compressing the files to be backed up with TAR as a compressed archive, and writing that data directly to a tape using shell scripts and programs like MT to control the tape drive. In such a case, restoring a single file would be very time consuming as the TAR archive would have to be decompressed before the individual file could be retrieved. On the other hand, the pitfalls of this approach have long been known and modern backup software uses sophisticated indices to know what data, is stored where, on which tapes. The court fails to distinguish between antiquated backup methods, and modern systems which can very easily restore single files.

The court further cites the lack of uniform standards in tape technology, which would be understandable had this case not arisen after the commercial introduction of the first generation of Linear Tape-Open (LTO-1) in 2000[34]. The industry, necessarily, has shifted towards open formats and continues to do so in all aspects of technology. That this drive for interoperability would also touch the backup tape market is expected. The other main tape formats DAT[9] , and DLT/SDLT[33] are also driven by standards, and

[9][38, at 319]

[10]Tape backups on unix based systems before the advent of robotic tape libraries required significant human intervention and manual tracking. Generally, individual file recovery was not possible without restoring an entire archive.

backward compatibility. It is unclear where the court came upon the idea tape storage ranged from the hundreds of Kb to several Gb when tape media available at the time was at 100 and above Gb capacity, and tape from ten years prior was at several Gb capacity.

Second, the court cites JBOD as type 3 – offline/storage archive – and references a document which is no longer accessible to justify this classification. Presumably drawing from the document, the court notes the primary difference between type 2 (near-line data) and type 3 (offline/storage archive) is that type 3 lacks, "the coordinated control of an intelligent disk subsystem". This is quite strange as the court uses JBOD and supplies the proper acronym expansion "Just a Bunch Of Disks". In modern usage, JBOD configurations are simply the combination of hard drives into a single volume without an underlying RAID setup – concatenating the storage if you will – and is generally achieved using a disk subsystem that also supports RAID setups. There is an archaic usage that goes by the same acronym but standing for "Just a Box Of Disks" which is referenced on Wikipedia and apparently entailed a computer with many disk drives mounted as separate volumes, but looking at periodicals around the time this case occurred makes it clear JBOD was used then[46] as it is now. In either case, such a storage system clearly falls under the court's type 1 classification.

Third, and most disturbing, is the court's type 5 classification. This classification is an example of a little technical knowledge inappropriately applied. The court equates fragmented files, damaged files, and erased files as similarly situated. In truth, damaged files and erased files are appropriately considered as inaccessible data types given they require significant effort to recover depending on the severity, and cannot be guaranteed as recoverable. The grouping of fragmented files with the latter two groups is perplexing, and indicates a fundamental ignorance of computer storage and file systems. The court relies on quotations from 'white papers' which are no longer accessible, but its interpretation of those papers indicates a belief that fragmented files require significant processing to access – an overstatement at best, and fundamental error at worst given the everyday nature of file fragmentation. The court could be referring to deleted files which are fragmented, but then would have to more specifically define deleted files or else the mention would be redundant.

The accessibility types promulgated by the court here are not as important to the Zubulake series as they are to the amended Federal Rules of Civil Procedure. We will see that the representation of a data source as accessible or inaccessible is significant for discovery under rule 26(b)(2). Using the flawed reasoning presented in the Zubulake type classifications, many types of storage can be represented as accessible or inaccessible depending on what features were emphasized.

The end result of Zubulake I was that UBS would conduct a search of its records stored on optical disks, and allow Laura to select 5 backup tapes which UBS would restore, search, and produce responsive emails

from. The court would then use the results of this in determining how to handle the remaining data.

1.2 Zubulake III

On July 24, 2003, the court ruled[11] on the motion to shift (share) the costs of restoring, searching, and producing relevant emails from UBS backup tapes. The court engages in a lengthy analysis applying its previously established seven factor test to evaluate the appropriateness of cost sharing.[12] Though the cost shifting itself is not relevant to the issues considered here, the court provides the results of the additional discovery ordered in Zubulake I.

An external vendor was selected by UBS to conduct the restoration and searching.[13][14] Searching was conducted by means of custom searches designed by the vendor. All of the relevant optical disks[15] – which recorded email going from the traders to an external source, or from an external source to the traders – were searched; less than twenty emails were produced from the optical disks. The five backup tapes Laura selected contained over 1,000 unique emails returned by the search, of which approximately 600 were relevant and produced to Laura.[16]

Laura identified sixty-eight emails of the 600 which she argued were relevant to her case, and the court supposed the results from the initial five backup restorations would be representative of the remaining seventy-seven tapes.[17] In particular, Laura pointed to several emails which she asserted contradicted testimony given in deposition by UBS employees or assertions made by UBS itself[18]. In addition, evidence was found among the restored emails that at least some emails were deleted despite UBS instructions to the contrary, and were now, presumably, only accessibly from the backup tapes. [19]

> In sum, hundreds of the e-mails produced from the five backup tapes were not previously produced, and so were only available from the tapes. The contents of these e-mails are also new. Although some of the substance is available from other sources (e.g., evidence of the sour re-

[11] [37]

[12] In order to determine whether cost-shifting is appropriate for the discovery of inaccessible data, "the following factors should be considered, weighted more-or-less in the following order"

[13] Pinkerton billed UBS 31.5 hours for its restoration services at an hourly rate of $245, six hours for the development, refinement and execution of a search script at $245 an hour,17 and 101.5 hours of "CPU Bench Utilization" time for use of Pinkerton's computer systems at a rate of $18.50 per hour.

[14] Pinkerton then performed a search for e-mails containing (in either the e-mail's text or its header information, such as the "subject" line) the terms "Laura", "Zubulake", or "LZ".

[15] fewer than twenty e-mails extracted from UBS's optical disk storage system.

[16] The searches yielded 1,541 e-mails,13 or 1,075 if duplicates are eliminated.14 Of these 1,541 e-mails, UBS deemed approximately 600 to be responsive to Zubulake's document request and they were produced.

[17] At oral argument, Zubulake presented the court with sixty-eight e-mails (of the 600 she received) that she claims are "highly relevant to the issues in this case" ... a review of these e-mails reveals that they are relevant. ... Presumably, these sixty-eight e-mails are reasonably representative of the seventy-seven backup tapes

[18] [37, at 285-86]

[19] [37, at 287]("For example, the e-mail from Chapin to Joy Kim instructing her on how to file a complaint against Zubulake61 was not saved, and it bears the subject line "UBS client attorney priviledge [sic] only," although no attorney is copied on the e-mail.62 This potentially useful e-mail was deleted and resided only on UBS's backup tapes.").

lationship between Chapin and Zubulake), a good deal of it is only found on the backup tapes (e.g., inconsistencies with UBS's EEOC filing and Chapin's deposition testimony). Moreover, an e-mail contains the precise words used by the author. Because of that, it is a particularly powerful form of proof at trial when offered as an admission of a party opponent.63 Zubulake v. UBS Warburg LLC, 216 F.R.D. 280 (S.D.N.Y. 2003)

The court concluded continued restoration of the backup tapes would potentially produce responsive information.[20] It shifted costs for the restoration between UBS and Laura at seventy-five percent and twenty-five percent respectively. Restoration would continue, and the results of which came to light in Zubulake IV.

1.3 Zubulake IV

On October 22, 2003 – approximately three months after the restoration effort began post-Zubulake III – the court ruled on a motion by Zubulake for sanctions against UBS due to UBS not preserving certain data . During the restoration resulting from Zubulake III[37] a number of monthly backup tapes were missing and a number of, "isolated e-mails created after UBS supposedly began retaining all relevant e-mails were deleted from UBS's system".[21] UBS filled much of the gap by locating weekly backup tapes for restoration, but the restoration was not complete.

Zubulake IV contains an important opinion from the court laying the parties' duty to preserve electronic information, and the penalties for failing to carry out that duty.[22] In her motion for sanctions, Zubulake sought, *inter alia*, an ADVERSE INFERENCE INSTRUCTION[23] against UBS for SPOLIATION[24]. An adverse inference is an "extreme sanction" because its *in terrorem* effect makes it nearly impossible for an affected party to prevail on the merits.[25] In seeking an adverse inference, the seeking party must establish three elements:

(1) that the party having control over the evidence had an obligation to preserve it at the time it was destroyed; (2) that the records were destroyed with a culpable state of mind and (3) that

[20][37, at 289]("...the possibility that the continued production will produce valuable new information-some cost-shifting is appropriate in this case, although UBS should pay the majority of the costs. There is plainly relevant evidence that is only available on UBS's backup tapes. At the same time, Zubulake has not been able to show that there is indispensable evidence on those backup tapes (although the fact that Chapin apparently deleted certain e-mails indicates that such evidence may exist).").

[21][39, at 215]

[22][39, at 214] ("This opinion addresses both the scope of a litigant's duty to preserve electronic documents and the consequences of a failure to preserve documents that fall within the scope of that duty.")

[23][39, at 219]("Zubulake asks that the jury in this case be instructed that it can infer from the fact that UBS destroyed certain evidence that the evidence, if available, would have been favorable to Zubulake and harmful to UBS.")

[24][39, at 216]("Spoliation is the destruction or significant alteration of evidence, or the failure to preserve property for another's use as evidence in pending or reasonably foreseeable litigation. The spoliation of evidence germane to proof of an issue at trial can support an inference that the evidence would have been unfavorable to the party responsible for its destruction.") [Internal citations and quotations omitted.]

[25][39, at 220]

> the destroyed evidence was relevant to the party's claim or defense such that a reasonable trier of fact could find that it would support that claim or defense. In this circuit, a culpable state of mind for purposes of a spoliation inference includes ordinary negligence.[26]

Further, relevance is inferred when evidence is destroyed in bad faith demonstrated *via* intentional or willful destruction rather than when the destruction is negligent where the seeking party must prove the destroyed evidence is relevant.[27]

1.3.1 Duty to Preserve

Parties in litigation are obligated to preserve evidence when they have notice the evidence is relevant to the litigation, or they should have known it might be relevant to future litigation.[28]The court determines the ceiling for UBS' duty to preserve was when Zubulake filed charges on August 16, 2001. It then examines whether the duty arose earlier based on Zubulake's argument that it should have arisen as early as April 2001. Starting in April 2001, internal emails concerning Zubulake included "Attorney Client Privilege" notations.[29] In a deposition from one of the key actors on the UBS side, it was admitted the possibility Zubulake would sue was "in the back of [his] head".[30] The court notes that there is a difference between one or two employees contemplating the potential for a suit and the relevant or involved employees doing so when determining if the firm as a whole comes under a preservation obligation.[31]

On the preservation duty question, the court concluded the relevant people at UBS anticipated the litigation in April 2001. The court then articulates, "[t]he duty to preserve attached at the time that litigation was reasonably anticipated."[32]

Once the duty to preserve exists, the question turns then to what the scope of preservation is.[33] The court quickly discards the notion that a corporation, recognizing potential litigation, must exhaustively preserve every shred of paper and byte of data as such a heroic requirement would cripple large corporations.[34] One could easily argue it would cripple small and medium entities just as easily as large corporations, and perhaps more so because smaller entities do not have the technical resources or personnel which large corporations do. Nonetheless, while herculean efforts are not required, parties anticipating or part of litigation must preserve

[26][39, at 220](Internal citations and quotations omitted.)
[27][39, at 220]
[28][39, at 216]
[29][39, at 216]
[30][39, 217]
[31][39, at 217]("Merely because one or two employees contemplate the possibility that a fellow employee might sue does not generally impose a firm-wide duty to preserve. But in this case, it appears that almost everyone associated with Zubulake recognized the possibility that she might sue.")
[32][39, at 217]
[33][39, at 217]
[34][39, at 17]

"unique, relevant evidence that might be useful to an adversary."[35]

In refining its articulation of these general standards, the court identifies a duty to preserve documents made by or for individuals likely to be discoverable with the duty extending especially to "key players" in the case.[36] It identifies the individuals whose backup tapes were lost as falling into that category. The court identifies that mirror-image copies created at the time the duty to preserve arises meets the duty requirement, but that there are other ways to achieve the same result of retaining all relevant documents.[37]

> The scope of a party's preservation obligation can be described as follows: Once a party reasonably anticipates litigation, it must suspend its routine document retention/destruction policy and put in place a "litigation hold" to ensure the preservation of relevant documents. As a general rule, that litigation hold does not apply to inaccessible backup tapes (e.g., those typically maintained solely for the purpose of disaster recovery), which may continue to be recycled on the schedule set forth in the company's policy. On the other hand, if backup tapes are accessible (i.e., actively used for information retrieval), then such tapes would likely be subject to the litigation hold. [39, at 218]

The court goes on to note an exception where a company can identify the backup tapes a particular employee's documents are located on. In such a case, backup tapes of "key players" should be preserved if not otherwise available.[38]

In this case, UBS employees did not comply with the previously created document retention directives. Had the policies been followed, three of the missing tapes would not have been lost.[39] Additionally, UBS did not directly order the preservation of backup tapes in Hong Kong until Zubulake made her discovery request.[40] The court notes UBS had and breached a duty to preserve the missing backup tapes.

1.3.2 Culpable State of Mind

The court begins its culpable state of mind analysis by noting any destruction[41] of documents for which there exists a duty to preserve is, at minimum, sufficient to meet the negligence standard.[42] Though unnecessary to satisfy the ordinary negligence standard, the court evaluates the level of fault for the different instances

[35][39, at 217, Quoting other cases]("While a litigant is under no duty to keep or retain every document in its possession ... it is under a duty to preserve what it knows, or reasonably should know, is relevant in the action, is reasonably calculated to lead to the discovery of admissible evidence, is reasonably likely to be requested during discovery and/ or is the subject of a pending discovery request.")

[36][39, at 217-18]

[37][39, at 218]

[38][39, at 218]

[39][39, at 219](noting the tapes should have been retained for three years.)

[40][39, at 219-20] In particular, the Hong Kong tape contained backups of Tong, the Human Resource employee responsible for Zubulake.

[41]Except for destruction caused by events outside the party's control such as a fire.

[42][39, at 220]

of missing backup tapes. It notes whether the duty to preserve extended to backup tapes was a grey area and that it was thusly unsurprising a company might think it had no duty to preserve all backup tapes even though litigation was anticipated.[43] The ordinary failure by UBS to preserve all potentially relevant backup tapes was, then, negligent rather than being grossly negligent or reckless.[44]

Turning to the Tong tape, the court finds failure to preserve exceeding mere negligence. Because Tong was the Human Resource employee responsible for overseeing Zubulake, was engaged in continuing correspondence with her, and because the tape covered the time period from after Zubulake filed her EEOC charges, UBS's notice of its duty to preserve was unquestionable.[45] It establishes the facts of the Tong tape's disposition as grossly negligent and perhaps reckless - exceeding even the ordinary negligence standard required for the Culpable State of Mind prong of the test.

1.3.3 Relevance

The first two prongs of the test being satisfied, the court turns to the final prong looking to relevance. Recall, the articulated standard for a culpable state of mind in this context; where the state of mind is intentional or willful, the relevance of the destroyed data is inferred, but where it is negligent the moving party must prove its relevance. In the prior analysis, the court found a gross negligence or perhaps reckless standard of culpability, but did not find willful or intentional conduct. Zubulake would have to demonstrate the destroyed evidence was relevant and would have been favorable to her to succeed in obtaining the adverse inference sanction.[46]

The court notes that in Zubulake I and III it found emails on the UBS backup tapes relevant to the case, but that in Zubulake III it specifically held the sixty-eight emails[47] produced to the court did not demonstrate the gender discrimination Zubulake alleged. It then observes, "[t]here is no reason to believe that the lost e-mails would be any more likely to support her claims."[48] The court reasons it is unlikely the missing backup tapes contain more relevant information than the remainder of the backup tapes because they cover a time period prior to Zubulake filing her EEOC charge. It notes the tape likeliest to contain relevant information is the Tong tape, but that most of the Tong emails are preserved elsewhere.

[43] [39, at 220]
[44] [39, at 220] (Zubulake argued the spoliation was intentional or grossly negligent.)
[45] [39, at 221]
[46] [39, at 221]("This is equally true in cases of gross negligence or recklessness; only in the case of willful spoliation is the spoliator's mental culpability itself evidence of the relevance of the documents destroyed.")
[47] These were the emails submitted by Zubulake as being most relevant among all emails produced.
[48] [39, at 221]

1.3.4 Zubulake IV Result

The court rules Zubulake did not demonstrate the lost tapes contain relevant information, and fails to satisfy the third prong. It denies Zubulake's request for an adverse inference.

In sum, UBS had a duty to preserve the backup tapes, and was somewhere between negligent and reckless in failing to preserve them. It was not, however, willful thus placing a burden on Zubulake to prove the missing tapes were relevant and would support her claims. Because she could not produce proof of supporting evidence contained on those tapes which would not be available elsewhere or which was not already produced, she failed the final prong.

Even though Zubulake was not able to meet the adverse inference burden, she was able to show e-mails were destroyed which should have been produced to her. The court allows Zubulake to re-depose certain witnesses at UBS's cost to inquire into the destruction and newly discovered emails. As we will see in Zubulake V, this will have tremendous implications for the discovery process with regard to electronic data.

1.4 Zubulake V

The fifth Zubulake decision came after the re-deposing of various defense witnesses as ordered in Zubulake IV. During the reconducted depositions, Zubulake became aware of additional deleted emails and unproduced emails still existing on UBS systems as well as evidence UBS employees deleted relevant emails (some of which were not recoverable).[49]

1.4.1 Willful Spoliation

Chapin is the UBS employee Zubulake alleges as the primary offender in her gender discrimination suit. It was identified, *via* an "oblique reference" in other places, that at least one email deleted by Chapin was non-recoverable. Chapin sent an email at 10:47 AM on September 21, 2001 requesting a document from Kim, another employee, quoting a conversation he overheard Zubulake having. Chapin then sent an email to the Human Resource personnel handling Zubulake (presumably Tong) and his boss complaining about Zubulake; the email contained Kim's supposedly verbatim recollection of Zubulake's conversation.[50] There is an email from Kim to Chapin at 11:19 AM on September 18 with the subject "2" - it contains a different supposedly verbatim quote from Zubulake than the one Chapin included in his email to his superiors / human resources. The missing email with the quote Chapin did use would have been sent between 10:58 AM and 11:19 AM with the subject "1".[51]

[49] [40, at 426]
[50] [40, at 427]
[51] [40, at footnote 29]

The first email with the quotation Chapin did use was deleted, was not recovered, and was presumed lost.[52] This was the tipping point of sorts. Because in civil discovery the requesting side must rely on the good faith efforts of the producing side, it becomes very difficult to show bad faith acts when dealing with massive amounts of data. In this case it took five Zubulake decisions before a single email was shown to be deleted intentionally, and be irrecoverable; the court correctly notes where there was one, there may be more.[53]

> Nonetheless, many backup tapes for the most relevant time periods are missing, including: Tong's tapes for June, July, August, and September of 2001; Hardisty's tapes for May, June, and August of 2001; Clarke and Vinay Datta's tapes for April and September 2001; and Chapin's tape for April 2001. Zubulake did not even learn that four of these tapes were missing until after Zubulake IV. Thus, it is impossible to know just how many relevant e-mails have been lost in their entirety.[54]

Even more damning, there is evidence that emails were deleted from active email stores after UBS's counsel sent out a litigation hold notice. This was determined, *exempli gratia*, by time lining an email present on Hardisty's August 2001 backup tape with when the litigation hold notice was sent, showing that it, obviously being present at least as late as August 31, must have been deleted after he was warned not to.[55] Other instances of this type of analysis are present in the decision.

1.4.2 Failure to Understand

I wish to highlight one specific failure in the case on the part of UBS's counsel. When UBS's counsel originally conducted their discovery efforts they made two significant errors regarding Tong, the Human Resources person overseeing the Zubulake issues. First, and obvious, they did not actually ask her to produce her files to counsel. Second, Tong indicated previously in two depositions that, "she kept a separate 'archive' file on her computer with documents pertaining to Zubulake".[56] UBS's counsel misunderstood the word archive to mean backup tape.

Reading between the lines of this section of the decision, it seems the "active data" at issue is an email archive. I would postulate this was a local email store where she was archiving pertinent emails. This would not be an unexpected occurrence, as it is very common in corporate environments for users to maintain local copies of emails to avoid exceeding server quotas either in terms of space or of email retention time limits. Considering the situation presented, where a Human Resources employee indicates they have an archive of

[52] [40, at 427]
[53] [40, at 427] ("Although Zubulake has only been able to present concrete evidence that this one e-mail was irretrievably lost, there may well be others.")
[54] [40, at 427]
[55] [40, at 428]
[56] [40, at 429]

their email - it should be unreasonable to assume this meant a backup tape in the first place to anyone with an inkling of technical knowledge. Why would the average user have any knowledge of the backup system to the extent that they would suggest their emails were on backup tape as opposed to indicating they were keeping a local copy within the normal meaning of the word archive? The court will later discuss this problem in the context of counsel's duties in discovery.

1.4.3 Counsel's Duties

In response to the quagmire created in Zubulake I-V, the court explicitly lays out counsel's duties in discovery. These duties have a particular slant toward electronically stored information because of its volatile nature and its continuing increase in both volume and importance.

It begins by noting discovery obligations do not end when a litigation hold is implemented. Rather, the hold is the beginning and counsel is obliged to oversee compliance therewith. The court stresses the need for oversight, monitoring, and communication in ensuring the party's compliance efforts are successful. In particular, to "ensure (1) that all relevant information (or at least all sources of relevant information) is discovered, (2) that relevant information is retained on a continuing basis; and (3) that relevant non-privileged material is produced to the opposing party."[57]

In implementing a litigation hold counsel must be certain the hold extends to all potentially relevant sources of information. In doing so, it is important counsel understands the client's technical infrastructure and data practices.[58] This internal investigation by counsel includes identifying and, "communicating with the 'key players' in the litigation, in order to understand how they stored information."[59]

The court also addresses situations where a large company or a suit with a large scope are involved.[60] The court suggests one possible solution to large data sets is series of keyword searches to identify potentially relevant information.[61] Paralleling what will be discussed in Part II, the court suggests this much broader *corpus* can then be reduced based on a negotiated search term list into a manageable set of documents for review and production.[62]

[57] [40, at 432]

[58] [40, at 432]("To do this, counsel must become fully familiar with her client's document retention policies, as well as the client's data retention architecture. This will invariably involve speaking with information technology personnel, who can explain system-wide backup procedures and the actual (as opposed to theoretical) implementation of the firm's recycling policy.")

[59] [40, 432] ("Unless counsel interviews each employee, it is impossible to determine whether all potential sources of information have been inspected.")

[60] [40, at 432] (noting, "counsel must be more creative."

[61] [40, at 432] ("It may be possible to run a systemwide keyword search; counsel could then preserve a copy of each 'hit.' Although this sounds burdensome, it need not be. Counsel does not have to review these documents, only see that they are retained. For example, counsel could create a broad list of search terms, run a search for a limited time frame, and then segregate responsive documents.")

[62] [40, at 432] ("When the opposing party propounds its document requests, the parties could negotiate a list of search terms to be used in identifying responsive documents, and counsel would only be obliged to review documents that came up as "hits" on the second, more restrictive search.")

Counsel's duties do not end after the initial hold is complete, but extends on a continuing basis to ensuring discovery obligations are met including by supplementing disclosures. The Court suggests the continuing duty implies a duty to ensure the information is not lost, and that obligations are ongoing. Turning to the question of what that continuing duty means, the Court articulates a balance between the party and counsel. It notes that a lawyer cannot monitor a client as if it were a parent and child, but rather there must be a reasonable standard where the client, at a certain point, bears the responsibility for preservation failures. This reasonable standard is balanced by counsel's superior knowledge of, "the contours of the preservation obligation" meaning a party cannot be reasonably expected to implement a litigation hold properly without counsel's active supervision.[63] Perhaps the court is hinting counsel should treat the party as a teenager - sometimes showing responsibility, but other times not taking out the trash?

It concludes by articulating at least three actions counsel should take in carrying out its duties. First, issuing a litigation hold either at the beginning of litigation or when it is reasonably anticipated. Second, communicating with key players in the litigation and clearly communicating the preservation duty to them (along with periodically reminding them of the duty.) Finally, instructing all individuals to produce electronic copies of their data while also ensuring backup media is retained and safely stored.[64] It is, however, not sufficient to simply notify the party of the litigation hold and expect it will be successful.[65]

1.4.3.1 Result

After Zubulake V, the Court concluded UBS acted willfully and granted the adverse inference against them.

> You have heard that UBS failed to produce some of the e-mails sent or received by UBS personnel in August and September 2001. Plaintiff has argued that this evidence was in defendants' control and would have proven facts material to the matter in controversy.
>
> If you find that UBS could have produced this evidence, and that the evidence was within its control, and that the evidence would have been material in deciding facts in dispute in this case, you are permitted, but not required, to infer that the evidence would have been unfavorable to UBS.
>
> In deciding whether to draw this inference, you should consider whether the evidence not produced would merely have duplicated other evidence already before you. You may also consider whether you are satisfied that UBS's failure to produce this information was reasonable. Again, any

[63] [40, at 433]
[64] [40, at 433-34]
[65] [40, at 432] ("In short, it is not sufficient to notify all employees of a litigation hold and expect that the party will then retain and produce all relevant information. Counsel must take affirmative steps to monitor compliance so that all sources of discoverable information are identified and searched.")

inference you decide to draw should be based on all of the facts and circumstances in this case.[66]

The jury ultimately decided for Zubulake awarding "$9.1 million in compensatory damages, and $20.2 million in punitive damages".[67]

1.5 Zubulake Conclusions

The Zubulake cases illustrate the murky waters inherent in discovery generally, but significantly enlarged by electronically stored information. The court's view is limited to what is brought to it, the requesting party is limited in what it may bring to the court by what the producing party will give it, and the producing party is limited by its knowledge of its client's systems. The court and requesting party rely on the assertions made by the producing party and on the idea they will carry out their duties in good faith. Overcoming this assumption to persuade the court's intervention was a long, hard road for Zubulake. In many ways, Zubulake I-IV seemed like Sisyphus pushing his rock - doomed to failure. It was only in Zubulake V where the egregious nature of the discovery abuses could finally be demonstrated that the myth parted ways from reality.

When these sort of watershed cases occur, the decisions reverberate through the legal sphere. Suddenly electronic discovery was not only front and center, but now the primary focus of everyone's attention. As we will see in subsequent chapters, the impact has been substantial and created a situation where more technical acumen is required to ensure compliance with obligations. The initial gold rush of vendors seeking to capitalize on a new market led the area into a confusing cacophony which it is only now emerging from as case law and standards develop.

Going forward, I will explore what those standards are and how we, as a field, can best approach the civil arena fully cognizant of the differences and pitfalls posed by it versus the more traditional arena of criminal cases digital forensics involves itself with.

[66] [40, at 439-40]
[67] [44]

Chapter 2

Electronic Discovery Evolves

2.1 Metropolitan Opera

"A lawsuit is supposed to be a search for the truth, [...] and the tools employed in that search are the rules of discovery."[1] Metropolitan Opera v Local 100[29] is case between an opera company and a restaurant-worker's union which contained a significant discovery dispute occurring contemporaneously with the Zubulake decisions. While Zubulake is the more famous of the cases and demonstrated the extreme difficulties inherent in finding fault with the black box of discovery, Metropolitan Opera demonstrates the other end of the spectrum where the discovery failings were clear and overwhelming.[2] Metropolitan Opera also clearly lays out principles about what delegation of the discovery process is allowed.

The first discovery foray in Metropolitan Opera was carried out with Joseph Lynett as council of record. Lynett provided copies of the first document request to Brooks Bitterman (the union's Research Director) and William Granfield (then Secretary/Treasurer, and President at the time of the decision); neither Bitterman nor Granfield is a lawyer.[3] The initial document production to the Opera consisted of leaflets and form letters, and was devoid of internal documents generally expected during discovery.[4] Lynett assured the court searches were conducted and all relevant documents produced, but in reality he delegated the duty to Bitterman and was not himself in a position to confirm its sufficiency or completeness.[5] Lynett's instructions to Bitterman were to instruct the Union staff to retain "Met-related" documents without defining what that

[1] [29, at 181] (quoting Miller v. Time-Warner Communications, Inc., 1999 WL 739528 (S.D.N.Y. Sept. 22, 1999).)

[2] [29, at 181]("It presented the unfortunate combination of lawyers who completely abdicated their responsibilities under the discovery rules and as officers of the court and clients who lied and, through omission and commission, failed to search for and produce documents and, indeed, destroyed evidence—all to the ultimate prejudice of the truth-seeking process. As confirmed by discovery into the Union's and its counsel's compliance with the Met's discovery requests, both the lawyers and the clients exhibited utter and complete disregard for the rules of the truth-seeking process in civil discovery.")

[3] [29, at 185]

[4] [29, at 185]

[5] [29, at 186] ("Despite Lynett's having been presented with a last clear chance to remedy the situation by Lans' calling attention to the failures of production on May 24, compliance discovery confirmed not only that these representations in open court were false but also that a thorough search, albeit on an expedited basis, had not been conducted and, therefore, that there was no basis whatsoever for the representation.")

meant. Bitterman, in turn, only alerted some Union staff. Lynett, Bitterman, and Granfield conducted unspecified searches of file cabinets, but did not confer amongst themselves to coordinate these efforts nor did Lynett clarify documents included drafts, differing copies, and electronically stored information.[6] The Opera sent a duplicate discovery request directly to Bitterman, subsequently resulting in additional documents being produced without explanation as to why they were not produced in the first instance.[7]

Michael Anderson and his associate Jennifer Matis replaced Lynett subsequently. Anderson maintained Bitterman was the logical custodian for the discovery records, but neither visited the physical office nor validated Bitterman's search practices or methodologies.[8] The Opera sent Anderson a second document request tailored to encompass the first request, but also to be more specific so as to leave no room for misinterpretation; additionally concerns were cited about the Union not producing entire categories of documents including, *inter alia,* all emails.[9]

The Union subsequently replaced Anderson and Matis with new attorneys from the Herrick Feinstein law firm. It emerged that Anderson had not instructed the Union to refrain from deleting computer files, and had not put a retention policy in place; additionally, no attorney conducted or supervised the document production efforts.[10] The Court directed Union that the discovery requests applied to ESI.[11] Anderson mistakenly thought emails were automatically preserved by the server and did not focus on email preservation in his efforts; the Union agreed to contact all of its ISPs to attempt retrieval of deleted electronic documents (though they only contacted some).[12] After discovering the servers only retained emails for 30 days, Anderson instructed Brooks Bitterman and other employees to make "all possible adjustments to save e-mail" or print out hard copies; Union's efforts at retrieving deleted email consisted of asking people who they had emailed to forward those emails back, but were otherwise unable to recover such.[13] Of the emails they did have, the handling of their production was a failure. In particular, Michelle Travis was deposed and stated she forgot about turning over emails she was instructed to, did not begin saving emails until late in the case, and otherwise did not save them unless she happened to print them out.[14]

There were a number of changes in counsel representing the Union, and ultimately five document requests frustrated by insufficient productions. In Granfield's deposition, he revealed the initial computer searches

[6] [29, at 185-86]
[7] [29, at 186]
[8] [29, at 186-87] ("Passing whether Bitterman was or was not a logical choice for custodian, compliance discovery confirmed that his supposed search was incomplete and haphazard and that no lawyer ever made inquiry of him about what he did.")
[9] [29, at 188]
[10] [29, at 190]
[11] [29, at 190] ("[...] expressly making the order applicable to all work done on computers and all information sent out or received through computers.")
[12] [29, at 190]
[13] [29, at 191] ("The Union has retrieved missing e-mails by contacting the known addressees. It has also contacted the internet service providers, who to date have advised that deleted e-mails cannot be recovered.")
[14] [29, at 194]

consisted of only some of the computers, and even then the search consisted of glancing at document titles in the "My Documents" folders on each with no searches on the internal text either by the built in Windows FIND tool or otherwise.[15] Beyond that search, Granfield also indicated documents were saved to diskette which were not produced, and from which documents were often deleted as it became full.[16]. Granfield also disclosed several computers were replaced with new systems (the Court notes perhaps coincidentally or perhaps not) two weeks after the Opera's counsel requested a forensic computer expert examine the systems; he also indicated the old systems were "still around".[17] The Court comments even if the statements that the systems were retained and not replaced because of the suggested forensic examination, replacing them without notice after various court orders and an announcement by the Opera that it might engage a forensic expert is a willful disregard of their discovery obligations.[18]

In Bitterman's deposition, he indicated he could not recall what he specifically did in the way of searching or when he did it. In Yen's deposition, who was an associate taking charge of the discovery toward the end of the timeline prior to this decision, she indicates various actions taken in conducting compliance discovery. The specifics of her attempts were lackluster, and the court finds her attitude typical of attorney conduct in the case noting that the Union's lawyers are inappropriately attempting to shift the burden of discovery to the other side rather than conducting their searches in good faith.[19] The Court then notes the failure of Yen's supervising Partner from Herrick Feinstein, Michael Moss, to actively supervise Yen and address opposing counsel's constant complaints of inadequate searches and production rises to willful misconduct; it further notes even if he did supervise Yen and acquiesced to her actions and omissions it was still a willful failure to engage in responsible discovery.[20]

2.1.1 Result and Conclusions

The Court inflicts the ultimate sanction and grants summary judgment in favor of the Opera against the Union.

[15] [29, at 209]
[16] [29, at 209-10]
[17] [29, at 210] ("At this deposition on March 25, it was learned that the three computers the Met had seen in the shared computer room at the time of the Met's walk-through on March 15 were brand new. When asked when the three computers that had been there for several years were replaced with these new ones, Granfield said, "About two weeks ago." (Id. at 408). Perhaps coincidentally, perhaps not, it was two weeks earlier, on March 11, in a teleconference, that Met counsel asked for permission to have a forensic computer expert examine defendants' computers.")
[18] [29, at 210] ("Even accepting that statement as true, for the Union and its counsel to permit the computers in use during the relevant period to be dismantled without notice to Met counsel in the face of a year of protest by Met counsel that electronic documents had not been retained or produced, court orders to preserve and retrieve electronic documents and the Met's announcement that it might engage a forensic computer expert is a willful disregard of their discovery obligations.")
[19] [29, at 214] ("Yen's comment typifies the lawyers' attitude toward discovery in this case; instead of shouldering the responsibility imposed by the Rules to conduct a good faith search for responsive materials—which would have yielded the obviously-responsive Housecalling Sheets—defense counsel seeks to impose on Met counsel the burden of identifying with specificity documents in the Union's files that are responsive and have not been produced. The Union's lawyers have it seriously backwards.")
[20] [29, at 215]

There are four take away points to consider in this case. The first is the attention the court pays to the entrusting of the discovery process and searches to a non-lawyer, here Bitterman. This concern will be echoed in cases discussed *infra*. Second, we see a failure to understand the underlying technical practices and architecture the Union operates its computer systems under; this parallels the "Archive" failure in the Zubulake decisions as discussed *supra* at 1.4.2 on page 19. Third, we see the Court take notice of the multiple failures by multiple actors, lawyer and not, to show what, how, or when they searched which is similar to the issues at stake in the case studies in Part II. Fourth, we see the Court hint at certain behavioral requirements once a party has announced they will seek a forensic expert to examine the computer systems; we might draw from this a higher duty of care implied or imposed once such a declaration is made.

2.2 2006 Changes to the Federal Rules of Civil Procedure

The discovery process in the federal court system is governed by the Federal Rules of Civil Procedure ("FRCP")[13]. In 2006, the rules were amended to explicitly include requirements for discovery of Electronically Stored Information or ESI. The rule amendments were adopted in the wake of cases such as Zubulake and Metropolitan Opera to modernize discovery in the face of new and expanding electronic data within the scope of discovery. The rules were amended again in 2007 primarily for stylistic purposes, and again in 2010 with some further clarifications, however the major ESI / eDiscovery changes were a result of the 2006 amendments creating a new era in discovery.

2.2.1 Rule 16

Rule 16 is a house keeping rule governing pretrial conferences, scheduling, and management. The Rule 16(b) changes bring emphasis to the potential need to address ESI early in the litigation. It provides for a scheduling conference after the Rule 26(f) conference reports are provided to the court. It also allows the parties' agreements from the Rule 26(f) conference to be entered as an order essentially binding them to the agreement for the purposes of the litigation including discovery agreements under 26(f) and agreements regarding "claw backs" under 26(b)(5) for privilege, work-product, or trial-presentation material.[21]

[21][13, (2006) Committee Notes Rule 16(b)] Rule 16(b) is also amended to include among the topics that may be addressed in the scheduling order any agreements that the parties reach to facilitate discovery by minimizing the risk of waiver of privilege or work-product protection.

2.2.2 Rule 26

In 2006, Rule 26 was amended adding 26(a)(1)(B)[22] including ESI in the list of information which must be disclosed, "without awaiting a discovery request." This section now appears in Rule 26(a)(1)(A)(ii) after the 2010 amendments.

Rule 26(b) covers the scope and limits of discovery, with rule 26(b)(2)(B) exempting parties from providing ESI where the source is identified as, "not reasonably accessible because of undue burden or cost."[23] This echoes the reasoning of the Zubulake I Court in setting out its five categories of accessibility as discussed at 1.1.2 on page 10. The committee notes for this section indicate the rule was designed keeping in mind the specific nature of a technology will affect how burdensome obtaining information from it in discovery will be.[24] As I will discuss in the case studies in Part II, the limitations of the specific system can have a large impact on the practicality and effectiveness of a given approach. When paired with the inaccessible/accessible limitations, this sets up a significant potential for conflict over whether a source is inaccessible.

The committee notes also indicate, "[a] party's identification of sources of electronically stored information as not reasonably accessible does not relieve the party of its common-law or statutory duties to preserve evidence."[25] The notes describe a system where the producing party may initially identify sources as inaccessible, but must still preserve them should the point be contested or the inaccessibility be overcome by a showing of good cause. They further provide for, when the parties cannot agree, that the issue of accessibility be settled by motion to compel or protective order, but, as is the theme in the amended discovery framework, the parties must confer and attempt to work out their problems before involving the court.

Rule 26(b)(5) recognizes the increased risks of waiving privilege due to the high volume ESI represents and the difficulty ensuring complete review prior to production. It allows the producing party, should they inadvertently produce such information to the opposing party, to attempt to "claw back" that information by notifying the other side of its assertion of a claim of privilege or trial-preparation material. As the committee notes discuss, the rule does not address whether privilege is waived by inadvertent production, but does provide the procedure for resolving such questions.[26] As we will see, *infra*, inadvertent production

[22][13, (2006) 26(a)(1)(B)](B) a copy of, or a description by category and location of, all documents, electronically stored information, and tangible things that are in the possession, custody, or control of the party and that the disclosing party may use to support its claims or defenses, unless solely for impeachment;

[23][13, (2006) 26(b)(2)(B)] (B) A party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost. On motion to compel discovery or for a protective order, the party from whom discovery is sought must show that the information is not reasonably accessible because of undue burden or cost. If that showing is made, the court may nonetheless order discovery from such sources if the requesting party shows good cause, considering the limitations of Rule 26(b)(2)(C). The court may specify conditions for the discovery.

[24][13, (2006) Committee note on 26(b)(2)] It is not possible to define in a rule the different types of technological features that may affect the burdens and costs of accessing electronically stored information. Information systems are designed to provide ready access to information used in regular ongoing activities. They also may be designed so as to provide ready access to information that is not regularly used. But a system may retain information on sources that are accessible only by incurring substantial burdens or costs. Subparagraph (B) is added to regulate discovery from such sources.

[25][13, (2006) Committee note 26(b)(2)]

[26][13, (2006) Committee notes 26(b)(5)] Rule 26(b)(5)(B) does not address whether the privilege or protection that is asserted

and privilege waivers are a growing problem.

Rule 26(f) in its current form, provides for planning the discovery process and mandates a conference between the parties in establishing the discovery plan. The plan requires the parties to state their views or proposals on a variety of matters including initial disclosures, subjects of the discovery, discovery deadlines or phases, ESI issues including how it should be produced, agreements on procedures for exerting privilege or trial preparation protection, limitations of discovery, and so forth.

2.2.2.1 Expert Testimony / 2010 Amendment

In 2010, Rule 26 was amended with a significant amount of attention to disclosures / discovery involving expert witnesses. The committee notes indicate this was intended to address a problem with the 1993 version of the rules allowing too broad of a discovery into expert communications with counsel.

> Many courts read the disclosure provision to authorize discovery of all communications between counsel and expert witnesses and all draft reports. The Committee has been told repeatedly that routine discovery into attorney-expert communications and draft reports has had undesirable effects. Costs have risen. Attorneys may employ two sets of experts — one for purposes of consultation and another to testify at trial — because disclosure of their collaborative interactions with expert consultants would reveal their most sensitive and confidential case analyses. At the same time, attorneys often feel compelled to adopt a guarded attitude toward their interaction with testifying experts that impedes effective communication, and experts adopt strategies that protect against discovery but also interfere with their work. [27]

Interestingly, the disclosure requirement under Rule 26(a)(2)(A) and the written report requirement of 26(a)(2)(B) may not apply to forensic experts employed during the discovery process. 26(a)(2)(A) requires a party disclose "the identity of any witness it may use at trial to present evidence under Federal Rule of Evidence 702, 703, or 705" and 26(a)(2)(B) requires the disclosure be accompanied by a written report if the witness is specially retained to provide expert testimony or regularly does so as the party's employee. Federal Rule of Evidence 702 states, "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education may testify thereto in the form of an opinion [...]".

The key is the 702 definition concerning the witness assisting the trier of fact to understand evidence or a fact in issue in the context of a trial, whereas discovery occurs before the trial and may not, in the case of

after production was waived by the production. The courts have developed principles to determine whether, and under what circumstances, waiver results from inadvertent production of privileged or protected information.

[27] [13, 2010 Commitee Notes Rule 26]

a jury trial or where discovery is handled by a magistrate judge or special master, be before the trier of fact at all. Forensic experts employed solely for the discovery process might then not fall under the definition depending on how "may use at trial" is interpreted.

2.2.3 Rule 33

Rule 33(d) provides that a party may point the interrogating party to records which may be reviewed, in sufficient detail to locate them, and give them a reasonable opportunity to examine/audit the records or make copies, in lieu of answering an interrogatory. This includes ESI. The limitation being the burden of obtaining the answer in such a way be substantially the same for either side doing it to prevent hostile discovery practices.

2.2.4 Rule 34

Rule 34 is fairly short and straight forward, but the 2006 committee notes are extensive and significantly longer. Rule 34 simply specifies the procedure for a requesting party to serve a request within the Rule 26(b) scope including, *inter alia*, requests to inspect, copy, test, or sample electronically stored information. Though a short rule, the commentary of Rule 34 summarizes the essence of what the 2006 rule changes were all about.

> Subdivision (a). As originally adopted, Rule 34 focused on discovery of "documents" and "things." In 1970, Rule 34(a) was amended to include discovery of data compilations, anticipating that the use of computerized information would increase. Since then, the growth in electronically stored information and in the variety of systems for creating and storing such information has been dramatic. Lawyers and judges interpreted the term "documents" to include electronically stored information *because it was obviously improper to allow a party to evade discovery obligations on the basis that the label had not kept pace with changes in information technology.* But it has become increasingly difficult to say that all forms of electronically stored information, many dynamic in nature, fit within the traditional concept of a "document." Electronically stored information may exist in dynamic databases and other forms far different from fixed expression on paper. Rule 34(a) is amended to confirm that discovery of electronically stored information stands on equal footing with discovery of paper documents. *The change clarifies that Rule 34 applies to information that is fixed in a tangible form and to information that is stored in a medium from which it can be retrieved and examined. At the same time, a Rule 34 request for production of "documents" should be understood to encompass, and the response should include,*

electronically stored information unless discovery in the action has clearly distinguished between electronically stored information and "documents."[28]

Rule 34, as it existed before 2006, contained essentially an antiquated holdover from a primarily paper oriented litigation paradigm. The courts, as they often do, stretched the definition of "document" over time to cover rapid changes in the technical reality - essentially applying the type of reasoning found in equity by noting it was "obviously improper". When ESI began looking like more than just digital analogs of traditional paper documents, *exempli gratia* .doc files, and became more complex abstract concepts, *exempli gratia* databases, the stretched definition of "document" was insufficient and required the FRCP be amended to clarify and end the practice of dealing with specific instances *ad hoc*. The bulk of the committee notes reaffirm, over and over, that ESI is to be interpreted broadly and expansively; essentially a hedge against future unforeseen changes and advances.

2.2.5 Rule 37

Rule 37(f) was added in the 2006 amendment, and is now numbered as rule 37(e) in the current format[29]. It provides protection from sanctions, essentially a safe harbor, for ESI lost due to "routine, good-faith operation of an electronic information system"[30] which would encompass things like automatic tape backup systems which overwrite old tapes in their normal operation, email systems with time limited retention, and the like. Within the safe harbor, however, lurks the specter of Metropolitan Opera and Zubulake - *id est,* the safe harbor only protects when no duty to preserve exists.[31]

2.3 Progeny Cases

The Zubulake decisions and Metropolitan Opera clearly show the need for the changes made in 2006 to the FRCP. The cases I examine in this section are the progeny of Metropolitan Opera and, in most cases, also of Zubulake. I choose to examine the Metropolitan Opera progeny in particular, because it represents the rarer instance of a full, manifest, and clear failure in discovery which we might term a boundary case for

[28] [13, (2006) Committee Notes Rule 34] (Emphasis Added)

[29] [13, Rule 37(e)] Electronically Stored Information. Absent exceptional circumstances, a court may not impose sanctions under these rules on a party for failing to provide electronically stored information lost as a result of the routine, good-faith operation of an electronic information system.

[30] [13, (2006) Rule 37(f) Committee Notes] Many steps essential to computer operation may alter or destroy information, for reasons that have nothing to do with how that information might relate to litigation. As a result, the ordinary operation of computer systems creates a risk that a party may lose potentially discoverable information without culpable conduct on its part. Under Rule 37(f), absent exceptional circumstances, sanctions cannot be imposed for loss of electronically stored information resulting from the routine, good-faith operation of an electronic information system.

[31] [13, (2006) Rule 37(f) Committee Notes] The good faith requirement of Rule 37(f) means that a party is not permitted to exploit the routine operation of an information system to thwart discovery obligations by allowing that operation to continue in order to destroy specific stored information that it is required to preserve.

the upper bounds of unacceptability. With such a boundary in mind, we can then work back through cases attempting to invoke its extreme standards to see where the results in those cases fall in an attempt to locate a lower bounds of what is acceptable.

Metropolitan Opera represented, essentially, the civil litigation equivalent of the death penalty for a case. Within the bounds of its upper bound extreme, and the lower bound of acceptable practice, the Zubulake decisions exist with the adverse inference standard. In the Metropolitan Opera progeny, we often also see the Court contemplate the Zubulake adverse inference tests which represent another bound lower than that represented by Metropolitan Opera, but greater than acceptable behavior. We may, in essence, be able to later formulate a range of infractions and penalties by examining what punishment various levels of abuse merit. A rubric of eDiscovery insufficiency.

2.3.1 American Friends of Yeshivat Ohr Yerushalayim v. United States

American Friends v United States[1] involved documents not produced to the IRS during discovery (despite court orders compelling its production) but for which American Friends sought to use as evidence. The Court sanctions American Friends by precluding it from using the documents in the case.[32] The Court relies on Metropolitan Opera [29] in concluding, "the knowledge of any of its agents as to the existence of the general ledger and tuition checks should be imputed to American Friends."[33] This case is also the progeny of the Zubulake[34] cases, and the Court relies on them in additionally considering that electronic files were not searched which presumably would have contained the missing ledger.[35]

The Court relies on Metropolitan Opera in considering the inadequate search echoing that case in noting no lawyer inquired into the searches performed by the non-lawyer.[36] It distills the Metropolitan Opera discussion into four categories of willful non-compliance.

> Metropolitan Opera is instructive here. In that case, the court found various instances of willful noncompliance with discovery obligations, including defense counsel's (1) "repeated[ ] represent[ations] to the Court that all documents responsive to the [plaintiff's] document requests had been produced when, in fact, a thorough search had never been made and counsel had no basis for so representing;" (2) failure to implement a document retention policy after the commencement of litigation; (3) failure to adequately explain to the non-lawyer in charge of production

[32] [1, at 2]
[33] [1, at 8]
[34] [40, at 431-32]
[35] [1, at 9] ("While Fishman vaguely asserts that he made repeated efforts to obtain documents from Strauss throughout the course of this litigation, Fishman admits that he never asked Strauss whether he had searched his electronic files, which presumably contained the general ledger")
[36] [1, at 13] ("The record thus reveals that Strauss' supposed search was incomplete and haphazard and that no lawyer ever made inquiry of him about what he did.") (Internal quotation omitted)

that discoverable documents included those in electronic form; and (4) failure to speak to and/or follow up with individuals who might have relevant documents.[37]

In granting the sanction, it ultimately concludes American Friends had clear obligations, and its failure was not due to factors beyond its control.[38]

2.3.2 Arista Records v. Usenet.com

Arista Records v Usenet.com[6] is a tangental progeny of Metropolitan Opera[39] but has a heavier influence from the Zubulake analysis for scope of preservation and duty to preserve. The issue here was the preservation and production of usage logs and other data which the defendants argued was transitory and that they were under no duty to preserve it, and alternatively it was technically infeasible to preserve the requested data for capacity reasons.[40] Regarding transitory data, the court notes the defendants' argument may have had merit prior to the request for the data's production, but after they had actual notice they gained an obligation to preserve it or negotiate a production of what data they could produce.[41] From a procedural stand point, the Court finds arguments regarding capacity issues unpersuasive in part because the defendants did not seek a protective order to address the burden they are alleging and they previously did produce some of the data being requested.[42] This potential to ruin a potential argument against burdensome production by failing to follow procedural rules or worse by demonstrating some technical ability to comply can be ruinous as it was here.

2.3.3 Convolve v. Compaq Computer

Convolve v. Compaq Computer[8] involved a patent and trade secrets dispute over Automated Acoustic Management (a technology which allows a hard drive to switch between quiet and performance mode). Convolve claimed the defendants were withholding discoverable information it was entitled to. The Court found some of the information, such as requests for bills of materials, disproportionally burdensome and/or relatively unhelpful as it did not specify which items in those documents actually used the AAM technology.[43] Convolve also argued Request for Proposal ("RFP")[44] and Request for Quote ("RFQ") documents were wrongfully withheld. Compaq had previously represented that it did not use RFPs/RFQs to the Special Master overseeing discovery; a subsequent deposition indicated this was erroneous, and that while it had only

[37] [1, at 13]
[38] [1, at 13]
[39] [6, at 429] (referring to the district court's discretion in sanctioning parties for discovery abuses.)
[40] [6, at 430-31]
[41] [6, at 431-32]
[42] [6, at 432]
[43] [8, at 167-68]
[44] Not to be confused with Request for Production.

recently formalized the use of RFPs/RFQs it previously used an informal equivalent system.[45] The Court notes that, while the misrepresentation was troubling, it does not imply the RFPs/RFQs should be produced due to the lack of information regarding the AAM technology in the represented items.[46] One RFQ, however, did call for drives with, "seek profiles switchable between quiet mode and performance mode."[47]

The Court ordered Compaq to search its systems for relevant RFPs/RFQs and submit detailed information on the steps it took to identify those documents.[48] This is a similar approach to what was ordered in my case studies which I discuss in Part II. Convolve also demanded direct access to Compaq's systems to perform their own searches as a sanction for the withholding; here the Court invokes Metropolitan Opera and sets an outer bounds on when discovery abuses would warrant such treatment.[49]The Court thus establishes widespread destruction or withholding of relevant information as the standard for circumventing normal discovery relying on Metropolitan Opera.[50]

2.3.4 De Espana v. American Bureau of Shipping

De Espana v. American Bureau of Shipping [10] is yet another progeny of both Zubulake and Metropolitan Opera. As of this writing, the case itself is not reported in F.Supp.2d, but is available in legal databases (2007 WL 1686327). The oil tanker Prestige was damaged and eventually sank near Spain; during the "casualty period" Spain responded to the cargo fuel leaked by the sinking tanker. Spain retained counsel and commenced suit against ABS. ABS sought, *inter alia*, production of email relevant to the casualty period; ABS later clarified it wanted records for approximately thirteen Spanish ministries, but later limited the scope to specific names and email addresses.[51] Spain refused on several grounds including that it was a fishing expedition[52], the computers were governed by Spanish privacy laws, and/or are protected by governmental privileges.

ABS moved to compel discovery, and succeeded. It further moved for dismissal under the Metropolitan

[45][8, at 168]

[46][8, at 168] ("Generally, RFPs and RFQs specified only the capacity and spin speed for the drives that were forecast to be required; there was no indication of any acoustic requirements.")

[47][8, at 168]

[48][8, at 168] ("Because this document is plainly relevant and because Compaq previously provided incorrect information about its RFPs and RFQs, it is important to ensure that its search has been comprehensive. Therefore, Compaq shall submit an affidavit setting forth in detail the steps taken to identify RFPs and RFQs, including those stored in electronic databases, which refer to the capability of switching between quiet mode and performance mode.")

[49][8, at 169] ("Had Convolve demonstrated widespread destruction or withholding of relevant information by Compaq, then sanctions, together with an order circumventing the normal process of discovery and allowing Convolve to access the data directly, might be appropriate.")

[50][8, at 169-70] (noting the only significant failure was the failure to produce RFPs/RFQs, "[b]ut that misstep is a far cry from the systematic abuse that served as the basis for sanctions in the cases cited by Convolve. See, e.g., Metropolitan Opera Association, Inc. v. Local 100, Hotel Employees & Restaurant Employees International Union, 212 F.R.D. 178, 181, 231 (S.D.N.Y.2003)")

[51][10, at 1]

[52]Fishing expedition as a colloquialism has been adopted by practitioners in civil law jurisdictions especially in continental Europe to refer to discovery beyond the scope of the limited discovery allowed in most civil law codes. As such, it may be approaching a "term of art".

Opera factors[53]. What is interesting here, is that in supporting its argument against the first factor (willfulness or bad faith) Spain points to both the quantity of evidence it produced, and its good faith efforts to engage forensic analysis after it was compelled.[54] The Court concludes Spain did not act in bad faith or with willfulness. Essentially, demonstrating attempted compliance using forensic analysis is being used as a defense against findings of bad faith or willfulness. This idea of "showing your work" is a common thread in these cases when problems arise - one might say, partial credit?

2.3.5 Richard Green (Fine Paintings) v. McClendon

In Richard Green v. McClendon[31], an art dealer sued a couple for breach of contract over failure to pay for a painting and promissory estoppel against the defendants' claim no contract existed. During discovery, Mrs. McClendon and her counsel represented they conducted the appropriate searches for documents in Mrs. McClendon's possession.[55] The plaintiff moved to compel production of other documents which the court granted; additionally the court ordered the defendants to, "certify the completeness of their responses".[56]

Mrs. McClendon produced an excel spreadsheet titled, "Fine Art, Miscellaneous Galleries" providing information about thirty-seven artworks including the one at issue in the suit. The plaintiffs requested additional information about the document; afterward Mrs. McClendon's counsel produced three additional versions of the spreadsheet.[57] The plaintiff then sought to have Mrs. McGlendon's computer undergo forensic examination, however the plaintiff later dropped the request when Mrs. McGlendon disclosed her computer had its operating system reinstalled by the 'son of a friend'.

> When Mrs. McClendon responded, however, she disclosed a fact that was previously unknown to the plaintiff and to the Court: in January 2009, "the son of a friend" who is "familiar with computers" reinstalled the operating system on Mrs. McClendon's computer. During this process, all of her files were transferred from her hard drive onto four compact discs ("CDs") so her computer no longer contains the original version of any of the information that was stored there prior to the reinstallation process.[58]

[53][10, at 3] ("As for the first inquiry into willfulness or bad faith, ABS maintains that Spain has acted willfully and in bad faith by continually claiming to have produced all responsive, non-privileged emails, when, however, this Court found that Spain had failed to issue a timely litigation hold.")

[54][10, at 3] ("In support of its argument that there is no evidence of bad faith or willfulness, Spain points to the massive record of evidence that it has produced from the casualty period, the evidence that email was not the primary means of communication among the main responders to the casualty, and the good faith efforts to engage in forensic searching of emails once ordered to compel.")

[55][31, at 287] ("Mrs. McClendon and her counsel repeatedly represented that they had conducted thorough searches for responsive documents and had produced everything in Mrs. McClendon's possession.")

[56][31, at 287]

[57][31, at 287]("Thereafter, on June 9, 2009, Mrs. McClendon's counsel provided what appear to be three additional electronic versions of the spreadsheet with partial electronic history for each. There are some clear differences between the initial spreadsheet provided to the plaintiff in hard-copy form and the additional electronic versions.") (Internal citations omitted.)

[58][31, at 287-88] (Internal citations omitted.)

The plaintiff withdrew the request because he contended it was, under the revised facts, a useless exercise; he sought instead, *inter alia*, an adverse inference against the defendants for their breach of discovery obligations. The court finds an obligation[59] to preserve existed and that Mrs. McClendon and Counsel were at least negligent[60], but finds any potentially lost documents do not meet the special relevance standard articulated in Zubulake[61].

The first take away here is the continued emphasis placed on properly conducted searches and properly implemented litigation holds. The second, and more subtle, is the strategic error the plaintiff made by withdrawing his request for forensic examination. A general reinstall of the operating system, even with drive format, is not going to obliterate all prior data - it is not, as many mistakenly believe, a wipe of all data on the system.[48] In fact, all he would have needed from a forensic examination was some text or information unfavorable to Mrs. McClendon to potentially overcome the special relevance requirement for an adverse inference. *Id est*, he jumped the gun by moving for the adverse inference under the mistaken assumption a forensic examination would be useless when in fact it was the only vehicle at his disposal to meet the requirements for obtaining an adverse inference.

[59] [31, at 289-90] ("This duty arose no later than October 3, 2008, when this lawsuit was filed; at that point, the defendant should have known that such information would be relevant or could lead to the discovery of admissible evidence. [...] There is thus no question that Mrs. McClendon was obligated to preserve the electronically-stored documents at issue.")

[60] [31, at 290] ("Here, Mrs. McClendon and her counsel were at least negligent in failing to implement a litigation hold, properly search for responsive documents, and supplement discovery responses in a timely and thorough manner. Indeed, the failure to implement a litigation hold is, by itself, considered grossly negligent behavior.")

[61] [31, at 291] ("There is no evidence, however, that any destroyed documents would have been unfavorable to Mrs. McClendon. In fact, it is uncertain whether the plaintiff has actually been deprived of any information, since all of the files previously contained on Mrs. McClendon's hard drive were purportedly transferred to the CDs that are now in counsel's possession.")

Chapter 3

Post-2006

In this chapter I will examine pertinent case law from the post-2006 changes to the FRCP. These cases represent how the courts are applying the ESI discovery rules, how the duties of counsel are interpreted, and what the penalties for breaching those duties are. In most of the cases I examine I generically refer to the Court, but in a select few I refer, specifically, to Judge Grimm. Judge Grimm authored numerous novel decisions over the past few years regarding electronic discovery issues so I wish to highlight which opinions are his in order to more closely examine them later.

3.1 Mancia v. Mayflower Textile Services Co.

Mancia v. Mayflower Textile Services Co. ("MANCIA") involves a dispute under the Fair Labor Standards Act involving allegations the defendants failed to pay overtime wages and made illegal deductions from wages. In the October 2008 decision[28], Magistrate Judge Paul W. Grimm provides a detailed analysis of discovery obligations under the FRCP bringing special attention to Rule 26(g)[1].[2] In what we will see continue to develop, Judge Grimm begins by noting the need to sanction rule violations to preserve the integrity of the discovery process.[3]

> If primary responsibility for conducting discovery is to continue to rest with the litigants, they must be obliged to *act responsibly and avoid abuse.* [...]
>
> Concern about discovery abuse has led to widespread recognition that there is a need for more aggressive judicial control and supervision. Sanctions to deter discovery abuse would be more effective if they were diligently applied "not merely to penalize those whose conduct may be

[1] [13, Rule 26(g)] (requiring all disclosures and discovery requests, responses, and objections be signed by at least one attorney of record or the party personally, and, by signing, certifying the document is consistent with the rules, not interposed for an improper purpose, and not unreasonable or unduly burdensome.)

[2] [28, at 357-61]

[3] [28, at 357] ("If a lawyer or party makes a Rule 26(g) certification that violates the rule, without substantial justification, the court (on motion, or sua sponte) must impose an appropriate sanction, which may include an order to pay reasonable expenses and attorney's fees, caused by the violation. Fed.R.Civ.P. 26(g)(3).")

deemed to warrant such a sanction, but to deter those who might be tempted to such conduct in the absence of such a deterrent."[4]

Judge Grimm then goes on to articulate several "take away points" or principles to learn from Rule 26(g) and the commentary. He notes first, the duty to behave responsibly in discovery and to act consistently with both the spirit and purpose of the rules is an affirmative duty.[5] Second, the court must impose sanctions for discovery violations both to penalize the violators and to deter others from failing to comply with their affirmative discovery duties.[6] Third, lawyers must avoid making discovery requests without considering the cost or burden those requests represent.[7] Fourth, and a corollary to the third, the rule seeks to end the "equally abusive practice" of reflexive or boilerplate objections to discovery requests not based on particular facts.[8]

With regard to the fourth principle, Judge Grimm notes, "[i]t would be difficult to dispute the notion that the very act of making such boilerplate objections is prima facie evidence of a Rule 26(g) violation, because if the lawyer had paused, made a reasonable inquiry, and discovered facts that demonstrated the burdensomeness or excessive cost of the discovery request, he or she should have disclosed them in the objection, as both Rule 33 and 34 responses must state objections with particularity, on pain of waiver."[9]

The undertone of this opinion touches on what the point of discovery is, and how it should be used. Discovery is about putting the evidence out in the open as efficiently and effectively as possible. Once the evidence is known, by both sides, the parties can evaluate their relative positions and may come to settle when the result is clear - thus eliminating the need for a time consuming and costly trial. On the other hand, when discovery is treated as a fencing game with requestor thrusting and responder parrying, the process frustrates the interests of justice[10]. A general or boiler plate objection to a request provides no information

[4][28, at 357] quoting [13, Rule 26(g) Committee Notes] (emphasis in original).

[5][28, at 357]

[6][28, at 358] ("As the Advisory Committee's Notes state, "Because of the asserted reluctance to impose sanctions on attorneys who abuse the discovery rules, Rule 26(g) makes explicit the authority judges now have to impose appropriate sanctions and requires them to use it. This authority derives from Rule 37, 28 U.S.C. § 1927, and the court's inherent authority."")

[7][28, at 358] ("the reality appears to be that with respect to certain discovery, principally interrogatories and document production requests, lawyers customarily serve requests that are far broader, more redundant and burdensome than necessary to obtain sufficient facts to enable them to resolve the case through motion, settlement or trial.")

[8][28, at 358] ("The rule and its commentary are starkly clear: an objection to requested discovery may not be made until after a lawyer has paused and consider[ed] whether, based on a reasonable inquiry, there is a factual basis [for the] ... objection. Yet, as in this case, boilerplate objections that a request for discovery is overboard and unduly burdensome, and not reasonably calculated to lead to the discovery of material admissible in evidence[...]") (internal citations and quotations omitted).

[9][28, at 359] citing [13, Rule 33(b)(4)] ("The grounds for objecting to an interrogatory must be stated with specificity. Any ground not stated in a timely objection is waived unless the court, for good cause, excuses the failure.")

[10][28, at 362-63] ("A lawyer who seeks excessive discovery given what is at stake in the litigation, or who makes boilerplate objections to discovery requests without particularizing their basis, or who is evasive or incomplete in responding to discovery, or pursues discovery in order to make the cost for his or her adversary so great that the case settles to avoid the transaction costs, or who delays the completion of discovery to prolong the litigation in order to achieve a tactical advantage, or who engages in any of the myriad forms of discovery abuse that are so commonplace is, as Professor Fuller observes, hindering the adjudication process, and making the task of the deciding tribunal not easier, but more difficult, and violating his or her duty of loyalty to the procedures and institutions the adversary system is intended to serve. Thus, rules of procedure, ethics and even statutes make clear that there are limits to how the adversary system may operate during discovery.") (internal citations, footnotes, and quotations omitted)

on what, if anything, is defective or burdensome about the request. A specific, factual objection provides necessary information for the requestor to reformulate the request into a better targeted or less burdensome one. Judge Grimm's continued emphasis on cooperation between the parties is grounded not only in the rules, but also in practical concerns for efficiency.

> Further, it is in the interests of each of the parties to engage in this process cooperatively. For the Defendants, doing so will almost certainly result in having to produce less discovery, at lower cost. For the Plaintiffs, cooperation will almost certainly result in getting helpful information more quickly, and both Plaintiffs and Defendants are better off if they can avoid the costs associated with the voluminous filings submitted to the court in connection with this dispute.[11]

In the instant case, the parties were to confer and discuss their discovery issues in hope of a cooperative solution.[12] After the conference, the parties still disagreed on a number of discovery issues; Judge Grimm, accepting the defense's representation certain documents did not exist, ruled the defendants did not need to produce further documents reflecting specific issues in the litigation.[13] The plaintiffs renewed their motion to compel citing three examples raising, "serious questions about the degree of candor prior representations made to the Court by counsel[...]" which caused Judge Grimm to reassess his ruling.[14] The motion to compel was granted, and the defendants were required to show cause why sanctions should not be imposed.

> In addition, they shall show cause why the Court should not order as a sanction that the Plaintiffs be permitted, at the expense of the Argo Defendants and their counsel, to have access to a mirror image, forensic copy of the electronically stored information of the Argo Defendants to be able to search for documents responsive to their document production requests. See Fed.R.Civ.P. 34(a)(1).[15]

Put simply, relying on Defense counsel's representations as truthful and in accord with their obligations, Judge Grimm initially ended that portion of the discovery. When shown evidence the Defense was untruthful or at least misrepresentative, he granted the motion to compel, then ordered the defense to show cause why he should not sanction them. Interestingly, and echoed in other cases examined in this part and in the Part II case studies, forensic examination appears to be the preferred solution when lawyers are shown to be

[11] [28, at 365]

[12] [28, at 356] ("I advised counsel that the dispute appeared to be one that could be resolved, or substantially minimized, by greater communication and cooperation between counsel and the parties, and provided detailed suggestions for counsel to follow at a meet and confer session.")

[13] [27, at 2] ("(1) the contract between the Argo Defendants and Defendant Mayflower; (2) hours worked by the Plaintiffs; (3) wages earned by the Plaintiffs; and (4) amounts paid by Defendant Mayflower to the Argo Defendants.")

[14] [27, at 3] ("Order, I denied the Plaintiffs' request for many categories of documents based on the factual assertion that they did not exist, or already had been fully produced. The Plaintiffs' Renewed Motion to Compel causes me to have concern that the representations that I relied on in my ruling were untrue.")

[15] [27, at 4]

acting in bad faith with regard to their discovery obligations. One might say digital forensics is a panacea for discovery shenanigans, albeit a potentially costly one if inexpertly applied

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

In Victor Stanley v. Creative Pipe[35], another decision by Judge Grimm, a dispute arose over a document production which included privileged ESI. At issue were two discrete *corpora*, the first contained 4.9 gigabytes of ESI in a "text-searchable" format and the second contained 33.7 gigabytes in a non-searchable format.[16] Originally, the production was produced in paper rather than in a digital format[17], but after the production sufficiency was challenged the court ordered the parties respective forensic experts to meet and confer.[18]

The experts agreed on a protocol and a list of search terms designed to locate responsive ESI within the *corpora*; the proposed search was not, however, designed to identify or eliminate potential privileged or work-product information from the result set.[19] The Defense's counsel notified the court that a complete, manual review would delay the production and come at "undue expense"; to address the articulated problem, the Defense gave its expert search terms to cull privileged or otherwise protected documents from the result set prior to production to the Plaintiff's counsel.[20]

Though Defendant's counsel originally requested a claw back agreement, it later dropped its request after the discovery deadline was extended and it estimated it could conduct a document by document review.[21] Subsequent to that review and production of the ESI *corpora* to the Plaintiffs, the Plaintiff's counsel discovered potentially privileged or work-product protected information in the production and notified the Defense of its inclusion. The Defense's counsel asserted privilege or work-product protection claims for each notification made by the Plaintiff's counsel, while later providing privilege logs.

Then, stage left, enters the conflict. Unsurprisingly, "[t]he 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"[22] setting up the conflict which will yield illuminating results for our

[16] [35, at 256]

[17] There continues to be a significant tendency for lawyers to print things out even where the original document is a native digital file such as .doc, .txt, .rtf, etc. In the modern context, post-2006, the tendency to produce as paper is further augmented by a fear of inadvertently producing privileged metadata, or of producing some other information not readily apparent from a cursory viewing.

[18] [35, at 254]("[...]meet and confer in an effort to identify a joint protocol to search and retrieve relevant ESI responsive to Plaintiff's Rule 34 requests. This was done and the joint protocol prepared.")

[19] [35, at 254] ("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.")

[20] [35, at 254-55] (The defense also acknowledged the potential for inadvertent disclosures due to the *corpora* size / document volume.)

[21] [35, at 255]("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.")

[22] [35, at 255]

purposes.

The Defense's position is that their expert conducted a privilege search using roughly seventy keywords decided on by various attorneys involved in the case, in order to segregate those documents as part of a preliminary stage in the review. The problem they discovered, as mentioned earlier, was the nature of the *corpus totum* being split into discrete searchable and non-searchable *corpora.* They applied the keyword list to the searchable *corpus* and conducted a page-by-page review of the non-searchable *corpus.* Later running into time constraints, they reviewed only the title of the documents in the non-searchable *corpus* only conducting a full page-by-page review when the title indicated potential privilege or work-product. The overall suggestion of the Defense's account is that they keyword searches culled the privilege documents from the searchable *corpus* and the manual review was a best effort under the circumstances to handle the non-searchable *corpus.*[23] The defense does not, however, explicitly identify where the 165 inadvertently produced documents were found, *id est*, in the searchable or non-searchable *corpus.*[24]

In contemplating the Defense's framing of the incident, Judge Grimm notes the Defendants were vague in describing the sufficiency and defensibility of their searches.[25] Speaking of the party and his attorneys, Judge Grimm notes nothing was 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"[26], and furthermore:

> [W]hile 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.[...]
>
> [T]he 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.[...]
>
> 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 overinclusive nor under-inclusive.[27]

[23] [35, at 256]

[24] [35, at 256] ("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.")

[25] [35, at 256] (noting the defense was vague in describing the keywords, how they were developed, how the search was conducted, and what quality assurance practices were in place.)

[26] [35, at 256]

[27] [35, at 256-57]

Judge Grimm thusly articulates important principles for defensibility and sufficiency. His comments highlight the need for careful consideration and rigorous practices in designing and executing ESI searches especially for privilege with necessary validation by way of sampling.

The Plaintiff's counsel had a far different view of the incident contending not only was it possible to perform the keyword searches on the text-searchable *corpus* possible, but that they conducted it with an undisclosed, but readily-available, search tool; further, the non-searchable corpus was primarily composed of image files and other like file types which are unlikely to be privileged or protected by the work-product doctrine.[28] They respond to the Defense's assertion of the difficulty in searching for the inadvertently disclosed documents by pointing out they were primarily located in the text-searchable corpus.[29]

In considering the standard used to determine whether a privilege waiver for inadvertent production occurred, Judge Grimm applies the intermediate test examining, "(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."[30] The Defendants bear the burden of proving reasonable conduct when assessing whether a waiver occurred; Judge Grimm notes their failure to provide pertinent information in meeting this burden including:

> [T]he 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; or whether they analyzed the results of the search to assess its reliability, appropriateness for the task, and the quality of its implementation.[31]

Judge Grimm further notes, "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*"[32]

In finding that the defense waived privilege, Judge Grimm discusses how the use of information retrieval methodologies require the "utmost care" in selection, design, and execution because the consequences of inadequacy can be dire. These considerations require, "advance planning by persons qualified to design

[28] [35, at 257] ("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.")

[29] [35, at 257] ("[T]he 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.at") (internal citations and quotations omitted).

[30] [35, at 259] (further noting the first factor strongly favors finding privilege was waived in the instant case.)

[31] [35, at 259-60](internal citations omitted)

[32] [35, at 260] (emphasis added).

effective search methodology" and the party using such must be prepared to explain and defend its use.[33]

Victor Stanley sets the tone for accountability in discovery involving ESI searches. As we will see in Part II, state rules are not so clearly developed yet but seem to be trending in the same direction of sufficiency and rigorous practice. The trend indicated here is a positive one for sound scientific practices, but calls for lawyers to engage experts early on in order to ensure the integrity of the process. As echoed in the AAFS 2010 business meeting for the Digital Media section, not all those currently holding themselves out as experts meet with our scientific standards; given the new ground being broken in the post-2006 period, this is not unexpected, but will have to be addressed in the near term.

3.3 Rhoads Industries, Inc. v. Building Materials Corp. of America

Rhoads Industries v. Building Materials Corp. of America[30] involves a suit where over 800 privileged emails were mistakenly produced by the plaintiff to the defendants. This case is the progeny of Victor Stanley[35] but also addresses the then recently enacted Rule 502 of the Federal Rules of Evidence ("FRE") [34], specifically addressing FRE 502(b)(2)[35]. Rule 502's approach is that of the "middle ground"[36] between conflicting precedents, and is similar to the "intermediate test"[37] employed by Judge Grimm in Victor Stanley. In addition to contemplating FRE 502, the Court also cites to Victor Stanley itself as an analogous fact pattern.[38]

In preparing for litigation, Rhoads engaged an IT consultant who tested and selected a discovery application for searching mail stores on its email system.[39] The court found the IT consultant reasonably believed the software would screen all privileged materials, and Rhoads' engagement[40] of such a consultant and screening tool showed its compliance with one of the Explanatory Notes[41] to Rule 502 concerning what

[33][35, at 261] (also referencing the Sedona Conference Best Practices)

[34][14, Rule 502] Attorney-Client Privilege and Work Product; Limitations on Waiver

[35][14, Rule 502(b)] Inadvertent Disclosure. When made in a federal proceeding or to a federal office or agency, the disclosure does not operate as a waiver in a federal or state proceeding if:
(1) the disclosure is inadvertent;
(2) the holder of the privilege or protection took reasonable steps to prevent disclosure; and
(3) the holder promptly took reasonable steps to rectify the error, including (if applicable) following Federal Rule of Civil Procedure 26 (b)(5)(B).

[36][14, Rule 502 Explanatory Note (2007)] ("The rule opts for the middle ground: inadvertent disclosure of protected communications or information in connection with a federal proceeding or to a federal office or agency does not constitute a waiver if the holder took reasonable steps to prevent disclosure and also promptly took reasonable steps to rectify the error. ")

[37][35, at 259]

[38][30, at 221] (summarizing the factors used in [35])

[39][30, at 221-22] The consultant selected Discovery Attender from Sherpa Software. The application is designed for searching Lotus/Domino databases for eDiscovery purposes.

[40][30, at 222] ("The fact that Rhoads retained a consultant who recommended and used a fairly sophisticated screening device shows that Rhoads' substantially complied with the following Explanatory Note to Rule 502")

[41][14, Rule 502 Explanatory Note (2007)] ("Depending on the circumstances, a party that uses advanced analytical software applications and linguistic tools in screening for privilege and work product may be found to have taken "reasonable steps" to

constitutes "Reasonable Steps" within the rule's meaning.

Rhoads's consultant ran a series of search terms provided by Rhoads's attorneys initially identifying 210,634 emails; subsequently, the consultant ran the search $*rhoadsinc * andeither * gowa*, *ballard*, or * cpmi*$[42] to cull out privileged emails. The privilege search identified 2,000 emails which were removed from the *corpus*; a second, revised, keyword search of the remaining 208,635 emails yielded 78,000 emails Rhoads believed were responsive and non-privileged.[43] The defendants notified Rhoads's counsel that certain produced documents appeared to be privileged; after review, 812 of the 78,000 emails were inadvertently produced privileged documents. Noting, for waivers of privilege, the party claiming the waiver bears the burden of proof, the court adopts a widely used five factor test in evaluating the waiver claim:

> (1) The reasonableness of the precautions taken to prevent inadvertent disclosure in view of the extent of the document production.
>
> (2) The number of inadvertent disclosures.
>
> (3) The extent of the disclosure.
>
> (4) Any delay and measures taken to rectify the disclosure.
>
> (5) Whether the overriding interests of justice would or would not be served by relieving the party of its errors.
>
> [30, at 219]

In considering the first factor, the court finds, *inter alia*, the acquisition and use of special software, the trial runs of the software conducted prior to its purchase, and the consultant's experience with Rhoads's computer systems weighing in favor of the plaintiffs. It likewise found the need for additional search terms, that the person reviewing documents for privilege had little experience and received little guidance from the supervising attorney, the limitations of the search to email addresses rather than the body, and the inadequate testing / validation of results / methodology weighing in favor of the defendants. The court analyzed all five factors, but in applying FRE 502 it found the heart of the dispute was whether or not Rhoads's actions were reasonable; thus, the first factor was most significant.[44] While the failure to adequately conduct its privilege

prevent inadvertent disclosure. The implementation of an efficient system of records management before litigation may also be relevant.")

[42] [30, footnote 5] ("Gowa Lincoln is Rhoads's law firm representing it in this case and other non-litigation matters. Ballard is another law firm that represents Rhoads. CPMI was retained by Rhoads as a non-testifying expert in this case.")

[43] [30, at 222]

[44] [30, 226-27] ("Once this lawsuit seeking millions of dollars in damages was filed, Rhoads was under an obligation to put adequate resources to the task of preparing the documents, which was completely within Rhoads's control. An understandable desire to minimize costs of litigation and to be frugal in spending a client's money cannot be an after-the-fact excuse for a failed screening of privileged documents, just as I refuse to use hindsight to criticize Rhoads for mistakes that were made but perhaps unforeseeable.")

review created a significant risk of waiver for Rhoads, its initial attempts at engaging technical expertise in the beginning ultimately tipped the decision in its favor leaving privilege intact.

The court here was somewhat critical of Judge Grimm's approach in Victor Stanley as relying overly much on perfect hindsight, but at the same time much of his analysis echoes Judge Grimm's. Incorporating the nationally applicable FRE 502 standard, this case illustrates further evolution in electronic discovery jurisprudence reaffirming the need for proper expertise and methodologies engaged early on in litigation.

3.4 Felman Production, Inc. v. Industrial Risk Insurers

Felman Production v. Industrial Risk Insurers[15], "is an insurance case, in which discovery has taken on a life of its own" and where privileged emails were inadvertently produced.[45] The Magistrate Judge overseeing discovery ruled the inadvertent disclosure by Felman Productions as waiving privilege, and the District Court considers the decision for clear error.[46] Feldman Production produced over one million pages of ESI documents marked confidential, admitted 30% of its production was irrelevant, and inadvertently produced nearly one thousand privileged emails in the process.[47]

In considering Felman's objections to the Magistrate's findings, the District Court applies both the FRE 502 and the Victor Stanley standards. Under the Victor Stanley balancing test, the court finds the large volume of privileged communications inadvertently disclosed and the "ridiculous" number of irrelevant items produced alone are sufficient to find Felman's pre-production actions unreasonable without analyzing the procedures themselves.[48]

The District Court specifically agrees with portions of the Magistrate's conclusions noting the failure to test keyword searches with sampling was imprudent and the quantity of inadvertent disclosures shows a lack of care in review.[49] In so doing, the District Court affirms the pre-production failure being sufficient to waive privilege, and no post-production actions can alter the result.

[45] [15, at 1]

[46] [15, at 3] (noting the standard of review for factual conclusions is highly deferential to the Magistrate Judge's determinations.)

[47] [15, at 1] (The Magistrate Judge ruled Felman waived privilege, "based upon its failure to take reasonable precautions to prevent inadvertent disclosure prior to production.")

[48] [15, at 3] ("To the contrary, as is suggested in the Victor Stanley balancing test, this Court finds that Magistrate Stanley was correct to judge the reasonableness of Felman's pre-production precautions and the question of waiver based on the results of the company's e-discovery.")

[49] [15, at 4] ("[...]the conclusions are consistent with the Victor Stanley test, in which the results of a producing party's inadvertence can and should be consider when determining whether waiver occurred, 250 F.R.D. at 259, and with Federal Rule of Evidence 502(b), under which an inadvertent party escapes waiver only if (1) the holder of the privilege took reasonable steps to prevent disclosure *and* (2) the holder promptly took reasonable steps to rectify the error.")

3.5 Harkabi v. SanDisk Corporation

In Harkabi v. SanDisk Corporation[17] the Plaintiffs moved for sanctions against Defendant SanDisk for failing to produce ESI, delayed production of ESI, and spoliation. The plaintiffs were the principal shareholders of MDRM, Inc. which developed technology for use with solid state drives; SanDisk acquired MDRM in 2004 and, as part of the sale, the Plaintiffs agreed to work for SanDisk after the acquisition.[50] SanDisk issued laptop computers and email accounts to the plaintiffs during their employment, and recovered the same when their employment ended; the Plaintiffs, in anticipation of a dispute, sent, via their counsel, a preservation notice. SanDisk's in-house council sent four memoranda instructing various actors within the company to preserve the Plaintiffs' laptops and other associated data.[51] In 2008, SanDisk changed to a new backup system for improved archival, and, soon after, the Plaintiffs' laptops were imaged and reassigned to other SanDisk employees.[52]

During discovery, SanDisk could not locate the laptop images and search terms were unsuccessful in locating the data the laptops contained on SanDisk's file servers.[53] SanDisk's counsel misrepresented the situation implying the laptops were recycled after 30 days under normal company policy rather than disclosing that the particular laptops at issue were stored for over a year then imaged, with the images subsequently lost.[54] SanDisk produced 1.4 million documents which it termed as "everything", then declined to produce the requested hard drive images (that it could not find) because, "all electronic documents from [the] hard drive[s] that are relevant to this dispute have already been produced."[55] Not surprisingly, the Plaintiffs found they were missing files they recalled being stored on their laptops.[56] In 2009, SanDisk acknowledged it could not locate the drive images despite "best efforts".

In reviewing what SanDisk did produce, the Plaintiffs found (1) fewer of their emails were produced than from other custodians, (2) emails the Plaintiffs were aware of were not produced, and (3) none of the emails from the Plaintiffs files were only between the two plaintiffs.

[50] [17, at 416]

[51] [17, at 416] ("Acting on those instructions, the laptops were placed in a secure storage area where they remained for more than a year.")

[52] [17, at 417] ("Thereafter, a helpdesk employee contacted SanDisk's Director of Information Security to ascertain whether the Harkabi and Elazar laptops could be reissued to other employees after imaging and preserving the data from their hard drives. That request was forwarded to SanDisk's in-house counsel and, according to the helpdesk employee, approved. Then, the Harkabi and Elazar laptops were imaged and the data saved on a SanDisk file server.") (Internal citations omitted.)

[53] [17, at 417] ("SanDisk began searching its file servers, but could not locate data from the Harkabi and Elazar laptop hard drives. In April 2009, SanDisk provided search term reports indicating that "Elazar" returned only five hits, and that no email was located for "Harkabi".") (Internal citations omitted)

[54] [17, at 417] ("At that time, SanDisk's counsel advised Plaintiffs' counsel that, when employees leave the company, their laptops typically are recycled 30 days later. However, SanDisk's counsel did not disclose that the Harkabi and Elazar laptops had been secured for a year and that efforts to locate the laptop data on SanDisk's servers had been unsuccessful.") (Internal citations omitted.)

[55] [17, at 417]

[56] [17, at 417] ("Despite considerable effort, Harkabi and Elazar could not find any of the materials they remembered being on their laptop hard drives—including meeting notes, calendar entries, and digital photographs of technical schematics drawn by Elazar on white boards —showing their involvement in developing the U3.")

> Reasoning from these data points, Harkabi and Elazar concluded that SanDisk had not actually produced any emails from their custodian files. They hypothesize that SanDisk cobbled together emails from other custodians and glossed over the fact that Harkabi's and Elazar's files were missing.[57]

SanDisk then disclosed it lost the Plaintiffs' email too.[58] They did, however, eventually find it in the backup tapes (after the plaintiffs moved for termination sanctions). [59]

Facts aside, Harkabi v. SanDisk applys the same general tests we observed in Zubulake and Metropolitan Opera house for adverse inference and termination sanctions respectively. The novelty of this case centers around five points of interest. First, despite the temporal distance from the aforementioned touch stone cases, the same types of bad practices keep cropping up with the same threat of sanctions, only now the courts have plenty of precedent to measure the failures against. Second, the court is openly willing to hold SanDisk to a higher standard because, "[i]ts size and cutting-edge technology raises an expectation of competence in maintaining its own electronic records" (though, apparently, an unmet expectation.)[60] Third, the court cognizes that, "a cascade of errors, each relatively minor, which aggregated to a significant discovery failure" which, at minimum, constitutes a sufficient negligence to proceed with the Zubulake style analysis for adverse inference.[61] Fourth, the court cites the upper and lower bounds in how harsh an adverse inference might be.

> In the most harsh formulations, a jury is instructed that certain facts are deemed admitted and must be accepted as true. The least harsh instruction permits, but does not require, a jury to presume that the lost evidence is both relevant and favorable to the innocent party.[17, at 420]

The terminating sanctions were not granted because they are a drastic remedy imposed in extreme cases, and here the Court found they were not warranted because there was no evidence SanDisk engaged in conduct such as intentionally destroying evidence.[62] The adverse inference was granted, with the harshness to be determined at trial. Finally, the court recognizes that SanDisk's misrepresentations were a significant impediment to the discovery process which, "[b]ut for Plaintiffs' forensic analysis and their counsel's persistence, those deficiencies may not have come to light"; the court subsequently imposes a $150,000 monetary sanction as compensation to the Plaintiffs.[63]

[57] [17, at 418](Internal citations omitted)
[58] [17, 418] ("SanDisk concedes that the native production did not include some Harkabi and Elazar emails because they were not preserved during transfer to the Evault system. The Evault transfer was implemented only for current employees. That Harkabi and Elazar were former employees subject to "Do–Not–Destroy" memoranda appears to have been ignored.")
[59] [17, 418] ("After briefing on this motion, SanDisk searched its backup tapes and began recovery of the missing emails.")
[60] [17, at 419]
[61] [17, at 419-20]
[62] [17, at 420] ("Terminating sanctions are a drastic remedy that should be imposed only in extreme circumstances, usually after consideration of alternative, less drastic sanctions. [...] Such an extreme sanction is not warranted here. There is no evidence that SanDisk engaged in egregious conduct such as the intentional destruction of evidence.")
[63] [17, at 421]

The repeated pattern of obfuscated discovery failings only being revealed due to persistence and forensic analysis points to a strong need for each side to be wary of the other side's claims and engage its own experts to keep their opponents honest. It also serves as a warning to us, to be alert for subtle clues hinting at such concealed failings.

3.6 Lee v. Max International

Lee v. Max International[25, 24, 26] is a novel case in that it rose to the appellate[64] level for consideration as to what discovery abuses rise to a sufficient failure to warrant a dismissal with prejudice or termination sanction. The Magistrate Judge considered[25] the Ehrenhaus Factors[65], and found the factors favored its decision to recommend the District Judge dismiss the case with prejudice. The Magistrate Judge came to this decision because of the plaintiffs' failure to comply with orders compelling discovery prejudicing the defendant, their delaying tactics interfering with the judicial process, the prior warning to the plaintiffs that continued non-compliance would result in the "harshest of sanctions", and lesser sanctions proved ineffective in forcing compliance.[25]

The District Judge adopted the Magistrate's recommendation, and dismissed the case with prejudice.[66] Reviewing the facts, the District Judge found the Magistrate correctly applied the Ehrenhaus factors by noting the plaintiff's repeated failures to comply with discovery orders.[67]

On appeal[26], the 10th Circuit affirmed the actions of the District Court holding three failures to comply with discovery obligations are sufficient to uphold the dismissal.[68] The Appellate Court then examined the facts relevant to the discovery failures. The defendants were unsatisfied with the plaintiffs' discovery productions and filed a motion to compel discovery; the Magistrate Judge granted the motion in October 2009 ordering production.[69] The plaintiffs produced a "trickle" of material and failed to produce many items the defendants requested and the court ordered produced.[70] The defendants moved for dismissal, but while the Magistrate Judge confirmed the plaintiffs disobeyed the discovery order, he gave the plaintiffs one last

[64][26, at 1320] ("Our district court colleagues live and breathe these problems; they have a strong situation sense about what is and isn't acceptable conduct; by contrast, we encounter these issues rarely and then only from a distance.")

[65][11] ("(1) the degree of actual prejudice to the defendant; (2) the amount of interference with the judicial process; (3) the culpability of the litigant; (4) whether the court warned the party in advance that dismissal of the action would be a likely sanction for noncompliance; and (5) the efficacy of lesser sanctions.") (internal quotations and citations omitted).

[66][24] ("The court agrees with this reasoning and hereby ADOPTS the magistrate judge's Recommendation and DISMISSES the plaintiffs' case with prejudice. IT IS SO ORDERED.")

[67][24] ("As described in the magistrate judge's order, the plaintiffs have repeatedly failed to follow the magistrate judge's discovery orders. [...] The magistrate judge twice ordered the plaintiff to produce various discovery, including state tax documentation. In the January 12, 2010 order, the magistrate judge warned that failure to comply would result in dismissal of the action. Yet, the plaintiff still failed to comply with all orders in that ruling.")

[68][26, at 1319] ("How many times can a litigant ignore his discovery obligations before his misconduct catches up with him? [...] We affirm. Our justice system has a strong preference for resolving cases on their merits whenever possible, but no one, we hold, should count on more than three chances to make good a discovery obligation. ")

[69][26, at 1319]

[70][26, at 1319]

chance to comply.[71] The plaintiffs certified to the court that all the requested documents were produced; the defendants did not find all of the requested documents and renewed its motion (the plaintiffs subsequently produced some of the missing documents).[72]

The appellate review for discovery challenges looks to whether the District Court abused its discretion in granting the sanction.[73] The 10th Circuit found the District Court was well within its discretion to dismiss a case when two orders compelling the same material are disobeyed when the materials are in its possession, custody, or control.[74] The 10th Circuit also noted such repeated failure is strong evidence of willfulness and bad faith[75] harkening back to the type of analysis used in Harkabi (see footnote 62 on page 46).

There is an additional subtle, but important, holding in this case. The plaintiffs argued they did not actually violate the second discovery order because, though they did not produce it when they submitted their declaration to the court that their production was complete, they did produce it before the February 26 deadline set by the Magistrate Judge.[76] The court does not accept their argument and finds a declaration under oath as to a production's completeness can be relied on by the opposing side:

> We disagree. Once the plaintiffs chose to declare— under penalty of perjury, no less—that their production of tax records was now compliant with the January 2010 order, the game was up. The court and defendants were entitled to take that sworn declaration to the bank, to rely upon it, to consider the matter closed. Yet, the plaintiffs produced the tax records only after Max uncovered the falsity of the declaration and only after Max was forced to file yet another motion concerning their production. None of this should've been necessary. And none of this, in any reasonable sense, demonstrates "compliance" with the January 2010 order. Discovery is not supposed to be a shell game, where the hidden ball is moved round and round and only revealed after so many false guesses are made and so much money is squandered. [77]

[71] [26, at 1319] ("Eventually, the magistrate judge in January 2010 confirmed that the plaintiffs had "blatant [ly]" and without apparent excuse flouted the October 2009 order. [...] [T]he court chose to give the plaintiffs one more chance to produce the requested documents. At the same time, the magistrate warned plaintiffs that "continued non-compliance will result in the harshest of sanctions.""")

[72] [26, at 1320] ("[T]he very next day Max sent a letter claiming that various materials still remained missing. Receiving no reply to its letter, on February 3 Max renewed its motion for sanctions. Two days after Max filed its motion, plaintiffs produced some of the missing records. Later in the month, the plaintiffs sent along yet more discovery materials.")

[73] [26, at 1320]("We view challenges to a district court's discovery sanctions order with a gimlet eye. We have said that district courts enjoy very broad discretion to use sanctions where necessary to insure ... that lawyers and parties ... fulfill their high duty to insure the expeditious and sound management of the preparation of cases for trial.") (internal citations and quotations omitted).

[74] [26, at 1320-21] ("We hold that the district court's considerable discretion in this arena easily embraces the right to dismiss or enter default judgment in a case under Rule 37(b) when a litigant has disobeyed two orders compelling production of the same discovery materials in its possession, custody, or control.")

[75] [26, at 1321] ("But a party's thrice repeated failure to produce materials that have always been and remain within its control is strong evidence of willfulness and bad faith, and in any event is easily fault enough, we hold, to warrant dismissal or default judgment.")

[76] [26, at 1322] ("They try to convince us that their false declaration shouldn't matter. The magistrate gave them, they note, until February 26 to comply with the January 2010 order. And though their January 25 production was incomplete and their declaration of compliance false, they eventually produced the requested tax records by February 26. And all's well that ends well, they say.")

[77] [26, at 1322]

3.6.1 Impact

In affirming the sanction, the 10th Circuit sets an important precedent binding on all its lower courts. As noted previously, discovery issues rarely reach the appellate level making this a significant decision. Though only binding on the 10th Circuit, its persuasive nature is felt in other circuits as well. It was cited, shortly after, by the 4th Circuit[78] in recognizing the deferential treatment of District Court sanctions in discovery for, *inter alia*, bad faith practices[79].

At the time of this writing, Lee v. Max International[80] has been cited in at least thirteen District Court opinions in ten different Districts, less than a year from being decided.[81]

3.7 Conclusions

In this chapter, I examined cases in the post-2006 era of electronic discovery. The trend is clear that courts will examine the qualifications and expertise of the individuals involved in discovery as well as the defensibility and sufficiency of their methods in determining whether discovery obligations have been met. We can see the Federal Rules of Evidence Rule 502 protections help streamline the process by mitigating the fear of inadvertent disclosure, but that those protections will not be construed as carte blanche to engage in unreasonable practices. We also saw that the dangers of not meeting discovery obligations can be dire.

As the case law develops, we are seeing a growing role for our field in ensuring proper practice on behalf of the side engaging us, and detecting improper practices on their opponents' side. We currently see the burden on the party crying foul to show such improprieties, but at the same time we also see some requirement developing for each side to justify their discovery methods. The continued expansion of our role is inevitable as the complexity of possible discoverable sources increases, *corpora* size grows, and discovery relies more and more on automated tools to handle the increasing amount of ESI.

[78] [36, at 7] ("We emphasize, however, that our review of the district court's determination is a deferential one, in recognition that "it is the district court judge who must administer (and endure)" the proceedings. Lee v. Max Int'l, LLC, 638 F.3d 1318, 1320 (10th Cir.2011); see also id. (advising appellate courts not "to draw from fresh springs of patience and forgiveness"). ")

[79] [36, at 12] ("[W]e find that the Appellants' bad faith throughout this litigation process was sufficiently egregious to justify the extraordinary sanctions imposed on them. Accordingly, we hold that the district court did not abuse its discretion. ")

[80] [26]

[81] E.D.Cal, D.Colo, D.Conn, D.D.C., S.D.Ind., D.Kan., S.D.Miss, D.N.M., N.D.Okla, W.D.Okla.

Chapter 4

The law now

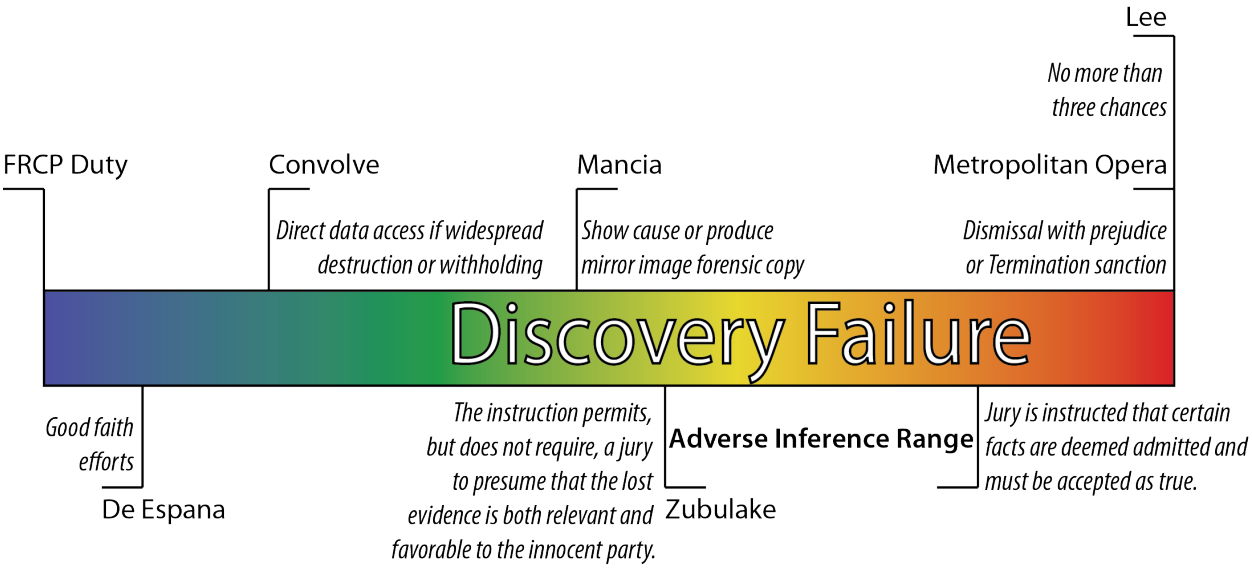


Figure 4.0.1: Discovery Failure Sanction Range

In Chapter 1, I discussed the Zubulake decisions which were a catalyst for changes to how electronic data is handled by the Federal Rules of Civil Procedure. Zubulake represents the legal equivalent of a discovery boogeyman - a reversal of fortune when new information is dramatically uncovered. In Chapter 2, I discussed Metropolitan Opera which occurred in the same timeframe as Zubulake, but where widespread incompetence represented a near comical failure by party and counsel alike. Taken together, they represent an unacceptable mishandling of Electronic Data in discovery which can easily occur if counsel is not sufficiently informed and vigilant in their discovery duties - a dawning realization of how deep the rabbit hole goes.

In 2006, the Federal Rules of Civil Procedure brought significant changes to address this new territory of amorphous electronic data which confounded the old methods of discovery created to handle physical documents. Much of the changes put an emphasis on affirmative duties had by counsel and imposed requirements for both sides to work together while seeking to curb potential abuses and run away costs. At the same time,

the progeny cases to both Zubulake and Metropolitan Opera show the courts willingness to use sanctions when the purposes of discovery are frustrated.

Following the rule changes, we see courts more aggressively discouraging discovery abuses and addressing failures with sanctions. In tandem with Magistrate and District Court opinions, a spattering of appellate decisions appear raising the specter of extreme sanctions such as dismissal with prejudice. Six years after the rule changes, much of the discovery process is still handled *ad hoc* rather than conforming to an agreed upon best practice. Some of this may vary with the relative sophistication of the parties and their counsel (litigation involving large technically oriented companies and law firms in primary markets tend to have better resources, and are better equipped for complex discovery than non-technical industries or firms in secondary or tertiary markets).

4.1 Forensic Computer Science / Digital Forensics in Civil Discovery

In criminal cases, the role of a forensic computer scientist is generally to preserve evidence, reconstruct events, and recover deleted data pertinent to the case. Such cases differ from civil discovery in that, unlike evidence seized by warrant in a criminal case, there is often limited or no access to the opposing party's systems pertinent data is drawn from. Because each side is responsible for producing their own data, we may be engaged in one ore more capacities. On one hand, we may be engaged by a firm to conduct the acquisition, preservation, and subsequently provide the data to be searched for discoverable material. On the other hand, we may be engaged to analyze what the opposing side has turned over to detect anomalies, missing data, or other discovery failures. In both cases, when a discovery dispute arises we may be called on to testify in support or in opposition to various discovery motions. The difference from criminal cases is that in civil discovery, why we are testifying is not always clear.

Figure 4.0.1 on the previous page illustrates a range of sanctions and failures derived from the cases I discuss in chapter 1-3. Where in a criminal case our testimony may concern whether or not an individual did something, in a civil discovery dispute often our testimony is about establishing the degree of the failure. It is not feasible to review every single byte of information a large company has at its disposal, and it is clear from both the rules and from the figure an exhaustive review is not the law. This shifts the question from did they / didn't they to was it or was it not reasonable in the context.

In determining reasonableness we may be called on to discuss specifics of the technology involved, estimates of how significant an undertaking a particular avenue may be, the costs involved in a particular ap-

proach, or the likelihood information may be obtained from a given source. As demonstrated in section 1.1.2 on page 10, previous analysis in other cases may be flawed or may no longer be true as technology changes. Concepts such as accessibility or inaccessibility of data is generally very fact specific. What discovery costs might be onerous in a suit for five figures in damages might be reasonable in a suit for seven.

4.1.1 The tipping points

In the preceding chapters I examined the events precipitating the 2006 rule changes, as well as post-change cases interpreting those changes. In the beginning of this chapter I noted the bulk of opinions concerning discovery practices are promulgated by Magistrate and District Court opinions. The dearth of case law emanating from appellate decisions is counterintuitive for a common law system where precedent tends to flow from above to harmonize practices below. This might come about because discovery is a procedural issue rather than a substantive one, and, as echoed in many of the discussed cases, falls within the court's inherent authority to manage its affairs.

Intuitively, it seems there must be contours to the courts' inherent authority in this context – certainly it would be dysfunctional if every court arbitrarily decreed their interpretation of discovery obligations on a whim, but at the same time too much interference from higher courts too far removed from the discovery process would prove stiffing to the lower courts' flexibility. The answer may lie with the nature of the rules themselves – namely that the Federal Rules of Civil Procedure, which governs discovery, is a Code. Serendipitously, being native to Louisiana and studying both civil and common law with a heavy emphasis on international and comparative law gives me certain insights to the approaches in both systems. Louisiana, in particular, is unique given its position as the sole Civil Law jurisdiction among its 49 Common Law sister states in a Federal system based on Common Law.

If we assume the FRCP, at least with respect to discovery, is being interpreted as codes are in Civilian legal systems, then how Louisiana interprets law should be informative. To that end, there are three cases of interest. The first, Ardoin v. Hartford Acc. & Indem. Co., is a well known case from the Louisiana Supreme Court in which Justice Dennis explicitly corrected a trend of Louisiana courts to adopt methods of interpretation from the Common Law. In the decision, the Louisiana Supreme Court explicitly affirms the nature of case law as a secondary source rather than primary – in effect, under the Civilian system case law has, theoretically, no more precedent than a treatise.

> In deciding the issue before us the lower courts did not follow the process of referring first to the code and other legislative sources but treated language from a judicial opinion as the primary source of law. This is an indication that the position of the decided case as an illustration of past

> experience and the theory of the individualization of decision have not been properly understood by our jurists in many instances. Therefore, it is important that we plainly state that, particularly in the changing field of delictual responsibility, the notion of Stare decisis, derived as it is from the common law, should not be thought controlling in this state.6 The case law is invaluable as previous interpretation of the broad standard of Article 2315, but it is nevertheless secondary information. [1]

The Louisiana treatment of judicial decisions as merely secondary to the primary authority of, *inter alia*, codes is unique in the United States. There are disputes as to whether this treatment is pure in practice, but at least conceptually Civilian methodologies are the rule. The second and third cases (Songbyrd, Inc. v. Bearsville Records, Inc. and Estate of Albritton v. United States) show examples of the State decision as applied in Federal courts which must apply Louisiana law.

> The basis of our jurisdiction, and that of the district court, to decide the instant case is diversity of citizenship, under which a federal court's obligation is to apply substantive state law. In Louisiana this obligation has special dimensions because of our unique Civilian tradition. We remain ever aware of the late Judge Rubin's caution to federal Erie courts applying Louisiana Civil law to steer clear of the common law principle of stare decisis and to apply instead the distinctly Civilian doctrine of jurisprudence constante: Because of the reviewing power of [Louisiana] appellate courts, the [Louisiana] trial judge may pay great respect to the decisions of these courts. He is not bound to do so, however, because the doctrine of stare decisis does not apply. Instead, each judge, trial and appellate, may consult the civil code and draw anew from its principles. Interpretation of the code and other sources of law is appropriate for each judge. The judge is guided much more by doctrine, as expounded in legal treatises by legal scholars, than by the decisions of colleagues.... Instead of stare decisis, the rule is one of deference to a series of decisions, jurisprudence constante.7 Emphatically elaborating on the proposition that Erie "does not command blind allegiance to [any] case on all fours with the case before the court,8 " now-Chief Judge Politz wrote that: If anything, this flexibility is even greater when a federal court sits as a Erie court applying the Louisiana civil law. In such cases, "the Erie obligation is to the [Civil] Code, the 'solemn expression of legislative will.' " Shelp, 333 F.2d at 439 (quoting the very first article of the Louisiana Civil Code). The Louisiana Supreme Court has taken great pains to "plainly state that ... the notion of stare decisis, derived as it is from the common law should not be thought controlling in this state. " Ardoin v. Hartford Acc. & Indem. Co., 360 So.2d

[1] Ardoin v. Hartford Acc. & Indem. Co., 360 So. 2d 1331, 1334 (La. 1978)

> 1331, 1334 (La.1978). While caselaw in the State of Louisiana is acknowledged as "invaluable as previous interpretation ..." [id. at 1335], it is nonetheless properly regarded as "secondary information." Id. at 1334.9[2]
>
> Erie of course applies with a civilian twist in Louisiana. It is axiomatic in Louisiana that courts must begin every legal analysis by examining the primary sources of law reflected in the positive codal and statutory expressions of legislative will. See,e.g., Prytania Park Hotel, Ltd. v. General Star Indemnity Co., 179 F.3d 169, 175 (5th Cir.1999). Jurisprudence, even jurisprudence that rises to the level of jurisprudence constante, constitutes only a secondary source of law. Id. On issues not yet squarely addressed by the Supreme Court of Louisiana, the federal court must rely primarily on the history and lineage of code provisions, their construction by commentators, and their place in the statutory framework. Delaune, 143 F.3d at 1002 & n.6; see also Songbyrd, Inc. v. Bearsville Records, Inc., 104 F.3d 773, 776-77 (5th Cir.1997); Green v. Walker, 910 F.2d 291, 293-94 (5th Cir.1990). Thus, a federal tax case in Louisiana may well be resolved primarily by reference to early nineteenth century French antecedents to the Louisiana Civil Code of 1825 together with French and Louisiana commentary. See,e.g., Delaune, 143 F.3d at 1002-1005 (the court of appeals resolved the tax issue by conducting an extensive analysis of the Louisiana code article's Code Napoléon antecedent and relevant scholarly comment). [3]

The corresponding concept of the Civilian system to the Common law's *stare decisis* is *jurisprudence constante*. Simply put, *jurisprudence constante* is, rather than a single precedent setting case, a line of cases with consistent outcomes. The courts, independently applying Civilian treatment, come to the same conclusion thus indicating a settled or clear point of law. Even *jurisprudence constante*, however, is not primary law in the sense a case decided under *stare decisis* is.

The trial courts have the inherent authority to manage their cases which includes managing discovery. The parts of the Federal Rules of Civil Procedure which govern discovery lay out both general principles and specific procedures. The inherent authority of the trial courts, therefore, must be bound by certain contours defined by these strictures. Combined with the tendency of discovery to be dealt with by magistrates whose decisions are of low binding potential, the discovery system behaves more like a Civilian system than a Common one. There is, however, *stare decisis* precedent as demonstrated by the appellate decision in Lee v. Max International, but, as noted, appellate decisions are rare in this context. This leads me to conclude the forbearance of the appellate courts and their deference to the trial courts and magistrates creates a situation where the certainty of precedent is eschewed in favor of the flexibility necessary in pre-trial discovery to

[2]Songbyrd, Inc. v. Bearsville Records, Inc., 104 F.3d 773, 776-77 (5th Cir. 1997)
[3]Estate of Albritton v. United States, CIV.A.98-859-A-M3, 2001 WL 1843696 (M.D. La. Dec. 10, 2001)

handle highly fact specific situations. This is, in the case of ESI, effectively required because the burdens and costs analysis must be conducted in the fact context of the instant case to determine reasonableness.

For these reasons, it is unlikely we will see significant jurisprudence developed at the appellate and supreme court levels regarding discovery limits in the near term. Discovery decisions by lower courts are given significant deference - usually for clear error only - because managing discovery is necessarily discretionary, fact specific, and vital to the district courts' ability to manage their cases efficiently. Too much precedent by higher courts would make the discovery process rigid, while not enough would make it non-uniform across districts. This may be viewed as a positive thing in that the constant state of change technology exists in does not function well with firm, long lasting decisions reaching into the future where the facts bringing them about are no longer congruous. There is no easy list to follow, but nor is there an antiquated and inflexible shackle of *stare decisis* to overcome. Because whether sanctions are imposed and what form they take is based on numerous factors, the best we can synthesize from the cases are general trends.

We see where minor or localized failures occur, the courts are likely to address them by compelling discovery, and addressing some failures, as they become more serious, with more rigorous forensic analysis. We might sum up this lower level by saying discovery was done wrong or badly, searches were not adequate, or data sources were not included. The sanctions are likely to address the failure by correcting it, and eventually producing what should have been given in the first place.

More serious are situations where pertinent data is lost and not recoverable. In this range we find the Zubulake case and the adverse inference standard. Detecting failures here can be difficult, expensive, and subtle. Because logging information is unlikely to be turned over in normal discovery, spoliation by an isolated actor within the company can easily go undetected. Zubulake itself would have turned out differently had textual references to the missing email not been found.

As will be discussed in Part II, sometimes failures can only be detected by familiarity with underlying systems. If one side submits a request for information to the other, and the only return is a set of documents, how can they be sure the other side's methods for responding to the request were reasonable and effective? Verification is difficult without access to search system used to fulfill the request, but often individual anomalies can be identified showing that, at least, something is amiss.

4.2 Future Trends

Data sets pertinent to litigation will increase as storage capacity increases. The increased datasets go hand in glove with more complex IT systems used in the corporate world. Shifts to cloud computing, distributed file systems, virtualization and similar technologies increasingly divorces the physical location and form of

data from the abstract presented to users. Lawyers are not experts in these systems, and even firms with skilled IT personnel often do not have sufficient institutional knowledge to deal with complex data sources while their counterparts in the client organization may have the institutional knowledge but not be aware of the legal nuances in discovery regarding sufficiency and reasonableness.

The role of digital forensics or forensic computer scientists is and will continue to be ensuring that what is claimed is what is true, identifying anomalies which may indicate discovery abuses or failures, and advising clients on the correct methodologies to employ.

Part II

Case Studies

Chapter 5

Case Study: Harris v BP

5.1 BELINDA HARRIS, ET AL. VERSUS BP AMERICA PRODUCTION COMPANY, ET AL

In December 2009, I was retained by the Talbot, Carmouche, and Marcello law firm ("TCM") as a technical advisor for their onsite discovery work at BP's Discontinued Projects Division office in Paleno Texas, and Conoco Phillips' corporate headquarters in Houston Texas. TCM obtained an order to compel discovery with significantly broad scope which allowed them, among other things, to run queries on databases, email storage, and other pertinent systems with the presence of their representative and a technical advisor. TCM suspected the defendants in HARRIS had not produced all responsive documents at their disposal, and were not using the most effective search capabilities at their disposal to locate responsive documents.

5.1.1 Order Compelling Discovery

The Court granted the Plaintiffs' Motion to Compel Discovery which I include here *infra* (exhibits omitted) in Appendix A on page 168. The order permitted the Plaintiffs' counsel to be present during the searches of BP's databases as well as to engage a technical expert to do the same.

5.1.2 British Petroleum Discontinued Projects Division

While at the British Petroleum Discontinued Projects Division ("BPDPD") office in Plano Texas, BP's representatives and counsel provided information on the systems they employed for locating boxes of documents which may be responsive. BP used a Concordance FYI v8 database system connecting to a series of Oracle databases at various locations throughout the world. Each database contained basic information about the contents of physical documents or boxes of physical documents including retention information, contents,

entry date, etc. Originally the fields were indicated to have a 512 character limit. This was not found to be entirely accurate, but close enough for analysis purposes.

BP's representatives indicated BP had no file servers, no central document management system for electronic documents, and that their concordance database was the only company wide searchable database. Prior to the discovery order permitting TCM to conduct discovery searches on site with BP's representatives, Duranda Smith, "the Manager of Discontinued Operations of Atlantic Richfield Company a wholly owned affiliate of BP America, Inc.", submitted affidavits detailing the quantity of search results returned from various queries. The search queries were ineffective in their design, which I will detail next, and likely a contributing factor to the broadness of the discovery order. Among other things, the order to compel ordered, "that plaintiff, with the assistance of its own IT representative and an IT representative of BP, is permitted to access and search any available searchable databases that may contain relevant discovery material. BP is ordered to provide information as to how its documents and searchable databases are organized and configured."

The Concordance system used by BP to index its archives of physical boxes is a boolean model or Black&White search system. The search functionality is quite robust, and it was clear from the affidavit listing search results BP's representative either was not adequately familiar with Concordance's operation or were deliberately using ineffective searches. I submitted an affidavit to this effect with my observations after my onsite visit with TCM's representative and, it is my understanding, BP was sanctioned for its conduct. The following is a list of example searches run by BP as listed in Duranda Smith's affidavit, the search results, and observations as to the search inadequacies

"North American Site Assessment Program" - returned no results. Search syntax ignores the possibility of plural Programs and derivative searches such as "North American Site Assessment", "Site Assessment Program", etc.

"Environmental Manuals" - returned 8 boxes. Search syntax ignores the possibility of singular "Environmental Manual"

"compilation of pit data" - returned no results. Search ignores alternative wordings such as "pit data compilation"

"general business policies and procedures" did not return any documents. It is not clear from the affidavit if they placed this in quotes, and if not how the 'and' between policies and procedures would be interpreted. The search also fails to account for singular versions and alternate wordings.

These are only an example of the queries run and are not exhaustive.

The Concordance system has robust search capabilities and expressive syntax supporting proximity searching using the 'near' keyword, wildcard expansion both fore and aft using '*', and complex boolean logic

evaluations. TCM's representative and I were able to construct significantly more complex and more effective queries as well as refine searches to exclude material BP's representatives objected to - such as excluding documents from specific people indicative of privilege, companies not acquired at the time the matter of the case occurred, or the like. The technical representative provided by BP was unfamiliar with anything but basic search syntax, the same was true for the other present BP representatives. As to the original assertion no file servers were in operation, an assertion I am incredulous of, I later noted BP's technical representative was saving the search results to a mapped network drive indicating the presence of at least one file server. When pressed again about file servers in a later session, BP's representatives did not recollect the previously asked question and provided response that there were no file servers

5.1.3 Conoco Phillips Corporate Headquarters ("CPCH")

CPCH used two primary document management systems. The first was LiveLink produced by OpenText, and the second was eSearch produced by IronMountain. The bulk of discovery I witnessed was conducted in LiveLink as the primary document management system. LiveLink contained information on physical boxes of documents as well as copies of native electronic documents. Each record had metadata associated with it for searching, but the system also indexed the native documents' content.

At CPCH I was present to advise an attorney for the plaintiff firm TCM, on the defense side was ConocoPhillips' inside counsel, outside counsel, several members of their litigation support staff, and an individual originally deemed to have sufficient system access to perform the required searches ("USER"). The initial course of action consisted of the support staff demonstrating the operation of the system and me consuming as much information as possible from the system documentation available online and through the user manuals. Additionally the TCM representative raised concerns regarding whether the level of access given to them through the provided search USER complied with the discovery order. For this case both sides were working off of the order to compel discovery granted against BP with the expectation an identical order could be easily obtained against Conoco anyway if they were not to cooperate fully.

In obtaining information on the operation of the LiveLink system and its access control configuration, Conoco's representatives indicated it was infeasible or impossible to produce a list of access missing from the USER in the LiveLink system. Symbolically, if we consider the universe of access {UNIVERSE}, with the user's access {USER} being a subset of {UNIVERSE}, we were interested in identifying {MISSING} as the relative complement of {USER} with respect to {UNIVERSE}. In identifying the sufficiency of the search results we needed to determine if {MISSING} contained documents or data subject to discovery or if the totality of {MISSING} were irrelevant, privileged, or otherwise non-responsive to the discovery requests.

Conoco's representatives indicated their system had over 500,000 groups defined for controlling access to LiveLink, no central list existed to explain what each group controlled access to, no individual administrator had total personal knowledge of the access groups, and standard practice for assigning access to users was on an ad-hock basis.

With assurance Conoco's staff was working to provide information on {MISSING} we conducted searches for later comparison. During this initial set of searches our protocol was to record the number of results returned and the query used, and save the search result pages from the web interface for later comparison. Later, Conoco's staff was able to create a tool for retrieving the information from the searches directly from the database for easier production. Once the results were obtained, they were reviewed by Conoco's counsel for privilege, confidentiality, or other relevance criteria before being turned over the the plaintiff's counsel as part of discovery.

While using the Livelink system for searches we noted discrepancies in the number of results returned for search queries. Discrepancies surfaced both with reruns of the same query and when shifting between one page of the search to the next, such as page 1 to 2, 2 to 3, 3 to 4, and so forth (see table 5.1.1 on the following page). This was our first indication the search engine used by Livelink did not have the same dependability in terms of exact searching as systems such as Concordance. From reviewing the documentation and after Conoco's representatives conferred with other Livelink support staff the consensus was reached that the number of results displayed were based on a timeout factor - e.g. it would gather as many results as it could before the time out and then display a total of "about x" results. As the system progressed to successive pages in the search the number would increase suggesting it was adding more results as we moved to the end of the list. For sanity checking purposes we also examined several documents to ascertain how the document was satisfying our search queries. We found that, especially for complex queries, the LiveLink search was operating with a relevance factor rather than only selecting documents which exactly satisfied our query; this grey searching is problematic in eDiscovery scenarios because it precludes us from excluding documents we would know to be privileged or irrelevant such as documents from counsel. We also noted through experimentation the Livelink system, likely due to its grey searching, is poor at evaluating overly complex search queries especially where multiple OR constructs (a OR b OR c OR d... OR n) are used with the prox[n,t/f] term[1] thus some searches had to be simplified and run in multiple equivalent searches.

Toward the end of the on site discovery Conoco's representatives were finally able to secure full administrative access for the searches in lieu of identifying {MISSING}. We reran several searches originally

[1]The Livelink documentation defines the PROX keyword as, "The PROX[n,B] keyword finds Livelink items that contain both specified query expressions within n words of each other. You set the value of n to a positive integer which specifies the required proximity (in words) of the two query expressions. The variable B represents an optional Boolean order flag that can be set to T (TRUE) or F (FALSE), depending on whether the order of the two specified query expressions is an important search component. By default, this variable is set to F (FALSE)."

Run	1	2	3	4
Results	694	779	819	821

Table 5.1.1: Changes in returned results for reruns of *grand chen**

A complex search query can be broken down into multiple simpler queries.

Complex Search	Simple Search
(A OR B OR C) PROX[10,F] X	A PROX[10,F] X
	B PROX[10,F] X
	C PROX[10,F] X

Figure 5.1.1: Equivalent simplified search example

executed under {USER} permissions to compare against the {UNIVERSE} results and found increases in search results from 175% to 2000%. The following chart illustrates the increases in search results returned. For confidentiality reasons the specific search terms have been made anonymous symbolically.

The nature of the increased result set is uncertain as the case settled shortly after, however the increase is sufficiently significant to indicate numerous relevant documents existed in other portions of the system to which {USER} did not provide access, thus putting them in the domain of {MISSING} which Conoco Phillips was unable to enumerate. Using {UNIVERSE} permissions eliminated the need to enumerate {MISSING} but logically we know some subset of {MISSING} controls access to the additional documents.

5.1.4 Harris v BP Conclusions

BP and Conoco Phillips are massive companies with significant technical infrastructure. It is not surprising the few representatives assigned to the case in question were not fully aware of every system's nuances. It was conveyed by Conoco Phillips' representatives that this case was the first ever in which plaintiff's counsel and representatives were allowed, by order, access to their facilities to search their systems. BP's representatives were significantly less informed than their Conoco Phillips' counterparts as to the design of their data storage and were less technically capable to search even what they were aware of. Conoco Phillips' representatives, while better informed and more willing to comply, were also faced with a significant challenge. From my conversations with the technical representatives provided both at BP and Conoco Phillips, it is my impression none of the provided technical representatives were computer scientists nor did they appear to be system administrators; the lack of qualified personnel assisting the defendants' attorneys is clearly a major factor in their systemic failure to conduct an effective search for responsive documents. Further, had I not been present TCM would have been significantly hampered in their ability to understand and utilize the systems placed at their disposal by the discovery order; acting alone they would have been forced to accept assumptions which would later prove inaccurate.

Search	{User}	{Universe}	Increase
(“dragon trail” prox[10,t] plant) and (envir* or contam* or groundwater or gw)	8	37	462.50%
(blanco prox[10,f] plant*) and (envir* or contam* or groundwater or gw)	17	302	1176.47%
(carney prox[10,f] plant*) and (envir* or contam* or groundwater or gw)	4	22	550.00%
(goldsby prox[10,f] plant*) and (envir* or contam* or groundwater or gw)	4	9	225.00%
(chittim prox[10,f] plant*) and (envir* or contam* or groundwater or gw)	4	7	175.00%
(maljamar prox[10,f] plant*) and (envir* or contam* or groundwater or gw)	20	400	2000.00%
grand chen*	821	2750	334.96%
spcc prox[20,f] plan*	502	3290	655.38%

Table 5.1.2: Increase in results between searches in {User} and in {Universe}

5.1.5 Generality

The case study I detailed previously illustrated a principle which can be generalized. I am reminded of a passage from Cyberpunk: outlaws and hackers on the computer frontier[45]:

> "Take a computer and put it in a bank vault with ten-foot-thick walls. Power it up with an independent source, with a second independent source for backup. Install a combination lock on the door, along with an electronic beam security system. Give one person access to the vault. Then give one more person access to that system and security is cut in half. With a second person in the picture, Susan said, she could play the two against each other. She could call posing as the secretary of one person, or as a technician in for repair at the request of the other. She could conjure dozens of ruses for using one set of human foibles against another. And the more people with access the better. In the military, hundreds of people have access. At corporations, thousands do. "I don’t care how many millions of dollars you spend on hardware," Susan would say. "If you don’t have the people trained properly I’m going to get in if I want to get in."

Susan expresses security reduction in terms of possible communications between two parties one of whom she could pretend to represent, this was her description of classic social engineering. The reduction in security she describes can be expressed as a directed K graph having $n(n-1)$ edges with each edge representing a potential social engineering attack. In effect, Susan’s explanation illustrates the lack of centralized knowledge as the number of actors within the organization increases; the lack of knowledge in a specific area by one of the actors then becomes an attack vector.

Thinking instead of an individual's knowledge of the total information storage and searching capability within an organization we can imagine the uncertainty which allowed an individual to be tricked instead representing lack of knowledge about some systems in the organization which may hold pertinent data; this constitutes a blind spot with regard to systems containing ESI. If the entire set of data storage systems in an organization is contained in the set {D}, and a given user u has access to {Du} a subset of {D}, then the uncertainty for that user would be expressed as the relative complement of {Du} with respect to {D}, which I will call {˜Du}. The greater the cardinality of {˜Du} the greater the uncertainty of the eDiscovery production's completeness. To minimize the uncertainty, eDiscovery must utilize a user set {U} to provide knowledge and access to systems which may contain responsive data such that we minimize {˜D{u}}.

There are different types of uncertainty which must be considered in this context. I have just identified knowledge of which systems exist, which I will term Awareness Uncertainty. Even in systems for which we are aware, however, we may not be able to eliminate all uncertainty. This comes about for access controlled systems in which knowledge of the system and access to the system does not imply total access to the system; without complete access to the system some uncertainty still exists. I term this Access Uncertainty. Finally, access to a given system does not imply understanding of how to effectively search the system or locate responsive data optimally. In cases where a user may be familiar with only a subset of the system's functionality or storage areas we can easily imagine a scenario where the user's inability to properly search for or locate the responsive data is the limiting factor to removing uncertainty. I term this Ability Uncertainty. Together these three areas, Awareness, Access, and Ability, form what I term 3A uncertainty.

In evaluating whether a given set of efforts exerted to locate and produce responsive documents and data is sufficient, we must examine the effort against the uncertainty in each area. The problem is having full knowledge of the optimal for each uncertainty area negates the need to compare in anything but a yes/no check. Instead we must establish heuristic judgement against what is reasonable to expect. This includes examining how knowledgeable our {U} is in total of the overall organization - ideally we will have overlap to give us an inkling we have captured the majority of knowledge available, the familiarity with the system the {U} has, and ensuring a proper level of access to the system is available when discovery searches are conducted. This problem is nontrivial and the heuristic is one provided by experience and familiarity with corporate structures, system administration, and other business processes. Evaluating how well a party has complied with their discovery obligations, or rather identifying if there are any deficiencies which a reasonably capable individual should have accounted for requires technical understanding way beyond lay men or "lay lawyers" as well as beyond the abilities of the average IT personnel.

5.1.6 Harris Conclusions

The BP and Conoco case study illustrates a level of ignorance which may exist in large corporate environments where no single individual is able to identify all systems containing discoverable ESI organization wide due either to the number of such systems, the presence of systems existing only at a local corporate level such as a department or location, or the complexity of data storage. The more complex the corporate entity the more difficult it is to eliminate or minimize uncertainty in the 3A areas. Because of the adversarial nature of litigation it is not prudent for parties to take the opposing side at their word without rigorous assurances and evidence they are meeting their discovery obligations. This necessary critical evaluation of the opposing side's compliance claims, the complexity inherent in large corporate IT infrastructure, and the difficulty in minimizing uncertainty in the 3A areas without the aid of experts hints at an expanding roll for digital forensics experts in civil litigation. The courts are also hinting through their evaluations of eDiscovery practices that design of searches, establishment of protocols, and ensuring efficacy in eDiscovery compliance may require experts to assist counsel.

The future of Digital Forensics on the civil side of litigation lies in a mediating or advisory role. To deal with this trend properly and prepare for the future we as researchers will need to expand our understanding of civil litigation especially eDiscovery and develop rigorous general methodologies to standardize the process. We need to develop a lexicon to describe systems which contain ESI in generic terms because court orders must use clear, specific language. We must study the corporate environment in order to develop rubrics for corporate scenarios to help evaluate how reasonable and effective an overall eDiscovery process is as well as to be capable of identifying what parts may be lacking. Most importantly, we must develop ways to minimize the invasive nature, cost, and productivity impact of eDiscovery. It is also vital we, as researchers, pay careful attention to the evolving legal landscape; without our input to guide the courts, decisions have and can continue to be rendered which, to put it politely, would not survive peer review.

Chapter 6

VPSB v. Louisiana Land and Exploration Company, et al

In this case study, I was engaged by the Plaintiffs shortly after they discovered, *inter alia*, that the DISCOVERY production from the Defense was incomplete and had subsequent to that successfully moved the Court to compel the Defense to produce the list of search queries used to identify responsive documents. The Plaintiffs contended they were not receiving all the documents they were entitled to. The Defense contended their production was complete and the issue being raised was an attempt to delay the trial. My role was to examine the Defense's search queries and render an opinion on whether or not the queries were properly designed to effectively identify responsive documents within the Defendant's "Times" database.

Like many legacy cases of this sort, there is a line of companies referred to as either successors or predecessors in interest. I will not delve into the issues raised in the case as to whether or not Chevron was or was not a successor in interest to the Louisiana Land and Exploration Company or any of the other entities. The pertinent issue is that the relevant document corpus is contained within the Chevron system; all other issues in the case are tangental to my discussion of the DISCOVERY.

6.1 Case Documents

Among my case studies, this case is unique due to the short turn around from my engagement to the Court ruling on the issue combined with the outcome. Additionally, I have the documents for the various stages illustrating the case's evolution through the motion to compel allowing us to see the inception of arguments to the final disposition of the situation's reality. The documents include the initial letter sent by the Plaintiffs' Counsel to the Judge (11.4) laying out their initial concerns, my affidavit (6.1.2) describing the search term sufficiency, the finalized articulation of the Plaintiffs' arguments (11.4), the Defense' counter-arguments (11.4), and finally the transcript from the trial itself (11.4).

The combined documents give a full view of how these disputes rise and are decided. These documents would generally be difficult to locate or acquire for scholars because they are part of the pre-trial phase of

the litigation and generally would not be available in the electronic indices used by lawyers for research.

6.1.1 Letter from Plaintiff's Counsel

The document included in Appendix B on page 173 is a letter from the Plaintiffs' counsel to Judge Winsberg updating him on their initial dissatisfaction with the situation. This letter was generated after I began my initial review of the Defendant's search term list. It is noteworthy because it provides an insight into a proto-argument. The letter contains the seeds of the arguments the Plaintiffs will eventually use in their Supplemental Brief in Support of Motion to Compel Access to Database. I would like to highlight one quote which summarizes nicely the difference between the old methods used in discovery, and modern methods.

> In a typical discovery production, some file clerk or records custodian who is familiar with the company's files works with counsel to gather all of the files that might contain responsive documents. Those files are then reviewed by counsel and any responsive documents are produced. The key to the reliability of the typical model is that the knowledgeable file clerk or record custodian can testify that, "Our review included all of our files that might have contained responsive documents." (p 1)

In the days where computer indexes were not widely used, lawyers relied on identifying human actors within an organization and tasking them with identifying all physical location of records. Once the totality of the potential document *corpus* was identified, the lawyers could then review the total *corpus* document by document, and produce documents responsive to discovery requests. In the modern sense, the corralling is being done by search queries. Thus, the search queries must be as effective as knowing permanent file locations was in the past. This problem is exacerbated by physical file locations existing, but no longer being the permanent or semi-permanent edifices they once were. *Id est* the document boxes are often managed by third parties swapping in and out of their custody, but managed *via* an electronic index.

6.1.2 Affidavit in Support of Plaintiff's Motion to Compel

My affidavit lays out the general problems raised in my analysis, but communicated in a way accessible to the Court and the parties. I will later more rigorously discuss search systems and how the different aspects impact searches. Herein, where I refer to a search as "black & white" or "positive / negative" I am referring to a system operating on the Boolean model of Information Retrieval.[43] (p 1) Where I refer to adequacy or inadequacy, sufficiency or insufficiency etc I am referring to the way the designed search conforms with or is likely to be effective in producing the substantial totality of responsive documents within the *corpus*.

(p 1) Where I evaluate the searches as indicative of the crafter being unfamiliar, unskilled, or careless in crafting effective searches, I am politely not including the third possibility that they are intentionally crafting ineffective searches. (p 2) I have not seen evidence of intentional obstruction in that sense, but it is always a possibility.

I talk about the need for real-time searches and cooperative refinement at various points. (p 2-3) I have previously presented my view of cooperative and collaborative work being the ideal for streamlining the problem of expanding document *corpora* sizes.[47] In my view, cooperative searching can quickly eliminate the need to argue about hypothetical and potentially excluded responsive documents. This is possible because revisions of searches done together will quickly show both sides if modifying a search to be broader or more constrained produces more responsive documents or more concise results respectively without shooting in the dark, so to speak.

15TH JUDICIAL COURT FOR THE PARISH OF VERMILION

STATE OF LOUISIANA

DOCKET NO. 82162 **DIVISION "D"**

STATE OF LOUISIANA and THE VERMILLION PARISH SCHOOL BOARD

VERSUS

THE LOUISIANA LAND AND EXPLORATION COMPANY, PEAK OPERATING CO., UNION OIL COMPANY OF CALIFORNIA, UNION EXPLORATION PARTNERS, LTD., CARROLLTON RESOURCES, L.L.C., and PHOENIX OIL & GAS CORPORATION

AFFIDAVIT OF BRIAN ROUX

Before me personally came and appeared:

BRIAN ROUX

who, after being duly sworn, did depose and say:

1.

I am an information technology specialist with extensive experience in database architecture and operations, as well as forensic investigation of IT systems. I am employed by VenueDocket in New Orleans and routinely provide consulting services to attorneys on electronic discovery matters.

2.

I have been provided with a list of search queries run for the case at hand to examine and advise the Talbot, Carmouche & Marcello law firm on the adequacy of those searches. I examined these terms under the assumption the unidentified search system used utilizes a strict boolean, "black & white", or "positive / negative" method. There are three areas I will discuss here (1) the inadequacy and inefficiency of using redundant and overly simplistic search queries, (2) the inadequacy and insufficiency of failing to use built in functions for searching alternate versions of words (plural AND singular for example) or of not including the alternate versions in the search where such systems do not support it automatically, and (3) the beneficial nature of cooperative real-time searching as a more efficient approach.

3.

First, the search queries used are simplistic and prone to duplication of documents produced. In particular, where multiple searches included the search terms ""Union Oil" AND Louisiana"

joined with another term, there is substantial likelihood documents would be produced multiple times. A more effective and efficient method of producing a concise set of searches would be to join the multiple search terms with **OR** constructions. For example:

"Union Oil" **AND** Louisiana **AND** NORM
"Union Oil" **AND** Louisiana **AND** "produced water"
"Union Oil" **AND** Louisiana **AND** audit

Could have better been searched as

"Union Oil" **AND** Louisiana **AND** (NORM **OR** "produced water" **OR** audit)

In such a case, documents which might deal with NORM as well as an audit report would be produced only once in the search result. The use of less effective searches, as present in the search list presented to me, is, in my opinion, indicative of either (1) the individual crafting the searches' lack of skill, knowledge, or familiarity with complex Boolean expressions, or (2) carelessness in crafting searches to be effective.

4.

Second, the search queries use limited variations of the terms they include. When singular or plural versions of a term are used arbitrarily, there is a greater risk of excluding pertinent documents from the search results out of proportion with the simplicity of including both versions of the search. In fact, many search systems support wild card character replacement and root expansion to automatically search for the alternative versions of the terms; some systems even provide the capacity for searching synonyms as well. The hypothetical example of a file named "Phase I & II Env. Assessments of SW Pit - VPSB Lease - East White Lake Field" would indeed be overlooked by defective searches failing to utilize the plural "Pits", account for the common abbreviation "Env", or to search for "Assessments" instead of "Survey".

5.

Third, real-time searches would be far more efficient and address the applicable time concerns. The benefit to strict boolean searches in a real time environment is the ability to develop and refine searches to exclude inappropriate information. In previous such scenarios in which I have consulted, this mechanism allowed plaintiff and defense representatives to quickly exclude irrelevant or privileged documents by expanding the search queries with NOT or negation expressions. In a real-time search, this process of refinement can take seconds, while in non-real-time environment

there is a delay that is orders of magnitude greater in communicating searches between both parties, and working out refinements before searching again. The nuances of the search system are often also opaque until the behavior can be observed if full documentation of the system is not available - an individual crafting the search who is not overly familiar with boolean logic may not be aware of better methods of expressing a search, or likewise observing quirks in the system may make the more effective search query less efficient due to the system's idiosyncrasies. In either case, real-time cooperative searching allows these problems to be addressed with speed and alacrity.

6.

The proper way to conduct a search for information to produce the most accurate results in the most efficient way is to do so methodically. The first cut must be necessarily coarse; likened to casting a large net. Without first capturing all the needed information, regardless of how much extraneous information is also caught up, the process of refining a search would be futile as the aim of the search has already escaped the net. Subsequent cuts in the data set become smaller and more refined either to exclude specific unwanted information, or to cast a smaller net within well stocked waters.

7.

Except with respect to Carrollton Resources' files, Unocal only ran searches for "East White Lake" and "VPSB" in combination with other terms such as "(sample or sampling)," "(study or studies)," "assessment," "survey" and "test." The results of these searches would have omitted files whose indexes included the terms "East White Lake" or "VPSB," but did not include one of the secondary terms listed above.

This affidavit is made on my personal knowledge, and I am competent to testify to the matters stated herein.

_____
Brian Roux

SWORN TO AND SUBSCRIBED
before me, Notary Public, this ____
day of October, 2010.

_____
NOTARY PUBLIC

_____
Printed Name of Notary

Notary No. _____

6.1.3 Plaintiffs' Brief in Support

The brief by the Plaintiffs is, like my affidavit *supra*, supportive of the Plaintiffs' motion to compel DISCOVERY. As we will see in the Defense' opposition included in Appendix C on page 177 this is the source of the "woefully inadequate" language misattributed to me. (p 1) The document is mostly inclusive of the letter on page 173, but contains refined arguments. The subtle but important legal issue expressed in the brief is the assertion the initial problems identified with the searches are sufficient to meet the legal standards to compel.

> These are just a few of the initial problems that plaintiffs see with the searches performed by Unocal. However, plaintiffs feel that these alone are enough to show that it is highly likely that important documents could be missing from Unocal's production in this case. (p 3)

As I will explore later, the Defense has a counter position on what is necessary to be shown, and that both are referring to what the legal standard for a *good faith showing* is under the Louisiana Code of Civil Procedure Article 1462.

The Plaintiffs mention the limitations imposed by the way the files are indexed as well as the search syntax of the system. (p 3) As explored throughout the case study chapters, this is a limitation common to oil companies where large *corpora* of physical records are manually indexed in text fields of limited length. The limited index information as compared to an OCRed[1] full text of the documents heightens the importance of properly crafted search queries.

6.1.4 Defendants' Opposition

The DEFENDANTS' MEMORANDUM IN OPPOSITION TO PLAINTIFFS' SUPPLEMENTAL BRIEF IN SUPPORT OF MOTION TO COMPEL ACCESS TO DATABASE was submitted to contest points and arguments raised both in the plaintiffs' corresponding brief and my affidavit. Like most legal writing, its content and language are influenced by the adversarial process and must be distilled through that lens to extract the meat of the arguments offered. It is very important to understand the underlying arguments at issue because those govern how we as a field must approach our work in cases such as this.

The brief starts out, "As predictably as snow in Alaska" and goes on to lament, "it is inconceivable that Plaintiffs will ever admit to satisfaction" with any list of search terms. (p 1) Though the defense accuses me of having, "never met a defense search term list that he did not characterize as 'woefully inadequate"'. (p 1) I must point out the "woefully" language is from the plaintiffs' brief, and that I simply described it, in less vitriolic terms, as a discussion of "inadequacy and insufficiency" of the methods used. (Affidavit, p 1)

[1] Optical Character Recognition

The defense phrases my role as being to, "generate discovery disputes so that Plaintiffs [...] can run amok in their opponents' computers and databases." (p 1) Mayhap the defense truly feels this way be it from emotional reaction or from *bona fides*, but misguided, belief. I cannot help but contrast the fanciful characterization of what the plaintiffs' are suggesting, what was done in the Harris v BP case, *supra*, and the idea of running amok in various computers and databases. Far from an uncontrolled and chaotic romp through their digital data-stores, the actions in that case consisted of us all sitting in a slightly warm conference room staring at a projector screen while methodically revising search queries with agreement between the parties. I believe the most exciting exchange there was my asking if anyone wanted some dried apricots[2].

It is important to understand what I testify to or submit affidavits about is but one facet of my work. In the totality of search queries I may be asked to review, only a subset is likely to be so inadequate or insufficient to merit challenging. This subset, the problematic instances, are what I discuss in the public documents. When discussing with my clients, the Plaintiffs' counsel in this matter, where I see problems and where I do not are both topics for discussion, but why would they ask me to submit an affidavit patting the defense on the head for the parts they did right? To the extent I seem only to criticize, it is because the criticism worthy elements are the only elements for which I would be asked to render an opinion at all.

6.1.4.1 Fairness

The defense paints a picture of an unfair situation by pointing to their inability to depose me, the lack of "actual" proof of existing documents not produced, and the lack of legal support for the plaintiffs' position. Addressing these in turn, I believe the defense is overstating both the strength of their objections and the burden of what is being asked for. (p 2) The court, as we will see in the transcripts in the section following, comes to the same conclusion.

First, did the defense have an opportunity to depose me? The defense filed its brief in opposition on October 29, 2010. The hearing in question where the matter was decided was on November 3, 2010. It is my understanding that at minimum, after the filing of their brief in opposition, the defense had an opportunity to depose me, but declined to. My records indicate me being told the defense did not seem like they wanted to depose me, but that they perhaps might bring their own expert to rebut. I can only assume, since I took the stand and was cross-examined, that the content of their questions then would have been the content of their questions at a deposition. It might be reasonable to assume that their lack of rebutting expert indicates either an inability to locate an expert willing to support the sufficiency of their searches, or that they did not believe the motion to compel had a significant chance to succeed.

[2]Whenever I travel by plane, I usually purchase dried apricots at the airport. We all have our eccentricities.

Second, can the "actual proof" requirement the defense argues for be the legal standard? In answer to this, referring to the transcripts in the section following, the court asks of the defense, "How can you prove what you don't know?". (Transcript, p 41) This question is the heart of the matter, and as we see the court is unmoved by the defense's *argumentum ad ignorantiam*. Because in civil discovery, unlike in a criminal case, the party is responsible for producing the evidence requested rather than having it seized by warrant, the opposing side can only argue from evidence something is missing or is likely to be missing based on circumstantial factors. The whole point of DISCOVERY is to discover relevant information. To require the thing to be discovered be already discovered in order to challenge its lack of production is contradictory to the aims of DISCOVERY itself. The defense takes issue with the plaintiffs constructing a hypothetical document which might be missed. (p 2) Is this unreasonable? I do not believe it necessary to construct a hypothetical document when the flaws in the search queries can be demonstrated more rigorously in a logical fashion; instead I see the "imaginary document" hypothetical to be simply an illustration and not a primary argument against the defense's practices. This, I think, is rather the defense attempting to place the evidentiary bar too high turning back the discovery clock to pre-Zubulake.

Third, does the law support the plaintiffs' proposed motion? The defense argues the remedy sought is beyond what the law is. (p 2-3) *Sed quid lex est*[3]? DISCOVERY is in the process of redefining itself as I have shown *supra*. Law on the federal level is clearly trending to holding parties accountable for their sufficient efforts, their own knowledge, and their good faith. Law at the state level often pays attention to Federal developments. I think it sound, *ergo*, to draw strength for the idea such measures are well within the court's discretion which will be discussed in the section following concerning Louisiana's Code of Civil Procedure articles 1461 and 1462. The defense argues the, "Plaintiffs and Mr. Roux would have the Court ignore the painstaking searches conducted by UNOCAL in this case in 2004, 2005 and 2006, before the onset of 'complex algorithms' or 'Boolean searches."'. *Cum scientia, res ipsa loquitur*[4]. They further talk about the "key determinant" never having been database searches, that tried and true methods of physical searching were used and are impervious to the "word games" employed by Plaintiffs; the defense, in essence, commits *argumentum ad antiquitam*.

The defense must advocate zealously for their client, so some leeway might be extended to them out of courtesy for their attempts to stave off the withering glare of scientific inquiry. We owe it to our field to approach the fight between plaintiff and defendant with a methodical and measured pace. The law is for the courts to decide, the scientific reality is for us. The march of progress must not be halted by the desire for a simpler time.

[3]But what is the law?

[4]With knowledge, the facts speak for themselves. A pun on the legal meaning of *res ipsa loquitur* to point out how ridiculous it is to say 2004-2006 was before the onset of complex algorithms or boolean searches.

6.1.4.2 Defendants' memorandum in opposition

The memorandum in opposition is included in Appendix D on page 183.

6.1.5 Court Transcript

A portion of the Court record containing my testimony on November 3, 2010 is contained in the Appendix on page 190. The transcript constitutes a primary source which might not otherwise be accessible, but is highly informative for the purposes of discussion. The discussion (6.2) refers to portions of the transcript by its internal page numbers. In the discussion I begin developing a formalized notation of the arguments.

6.2 Analysis

6.2.0.1 Framing the problem

The concept of boolean logic in a formal sense is very often outside the conscious cognizance of practicing attorneys despite their daily use of it. Lawyers in the current era utilize searches to comb through vast document *corpora* in search of data relevant to the discovery process in a specific case instance. In my testimony *supra* (p 15-16) I give a brief, and simplistic definition of boolean logic to establish the parameters of my inquiry into the discovery search queries at issue. One of the significant disputes here was the selection of search terms, which are the boolean components in the query. The dispute over search terms is very important because it is the determining factor over whether a document is produced or ignored. *Exempli gratia*, a search for "salt water" may vary with regard to the same document between two systems due to the inherent system idiosyncrasies. As we can see in the table below, the way a system matches can directly impact the search sufficiency especially where the ordinary human observer might assume the same result should occur for either search in either system. As I explain (p 16-17) this is also a problem for singular versus plural words, and instances where words may have a common abbreviation. [5]

Search Term	saltwater	salt water
saltwater	T	F
salt water	F	T

(a) Evaluation in a strict matching system

Search Term	saltwater	salt water
saltwater	T	F
salt water	T	T

(b) Evaluation in a partial or substring matching system

Table 6.2.1: Boolean evaluations in different matching systems.

As demonstrated *a posteriori* in the Harris v BP case study and shown in the table *supra*, the effectiveness

[5] *Exempli gratia*, "Saltwater" is often abbreviated "sw" in the records at issue here.

of a specific query will vary in search systems based on the type of search being conducted. We might think of the DISCOVERY process as the query Q_i in the specific system S_j operating on a universe of documents $\{Corpus\}$ in terms of evaluating $S_j(Q_i, \{Corpus\}) = \{Results\}$. The issue being alleged here is that[6] Q_∂[7] and Q_π[8] when evaluated resulted in $S_j(Q_\partial, \{Universe\}) \neq S_j(Q_\pi, \{Universe\})$, and also that where $\{Documents\} \in \{Results\}_\pi$ and $\{Documents\} \notin \{Results\}_\partial$ the $\{Documents\}$ are DISCOVERABLE[9].

There are further considerations, however, which impact how the evaluation might occur. As I noted, the search operates on a set of documents. Assuming the set being operated on is the $\{Universe\}$ can create problems when in actuality the evaluation occurs in a subset $\{$USER$\}$ constrained by authorization within the computer system. This problem is what occurred at Conoco Philips in the Harris v BP case, and is summarized in my testimony here (p 20-21). In further developing a formulaic representation of what the issue is, I must introduce a new function representing what documents are accessible to the user.

Assume, then, a function $A_j(U, \{Corpus\})$ where A_j is the access control system corresponding to system S_j and where $U \in \{Users\}$ represents a user within the system. It follows then that $A_j(U, \{Corpus\}) = \{Accessible\}$, that $\{Accessible\} \subset \{Corpus\}$, and that $\{Accessible\} \neq \{Corpus\}$ except where U is a "super user" with unlimited access to the system from which it follows that $\{Accessible\} \equiv \{Corpus\}$. We can finally represent the act of searching as a given query, in a given system, run by a given user as the following:

$$S_j(Q_i, A_j(U, \{Corpus\})) = \{Result\}$$

Figure 6.2.1: Functional representation of an act of searching in DISCOVERY

6.2.0.2 Handling results

Having defined the abstract concept of "searching" and reduced it to functional notation, I turn now to how $\{Results\}$ should be handled. In my testimony, the defense asks a series of questions about duplicative results. (p 21-22) It suggests, initially, that duplicative results would result in a duplication of effort in reviewing the documents in $\{Results\}$. My response, in essence, is an observation of algorithmic efficiency. In this case, as in Harris v BP, $\{Corpus\}$ and thus $\{Results\}$ do not contain documents themselves. Instead, the $\{Results\}$ contains indexes with short, human generated, descriptions of a box of documents. It is only necessary, *ergo*, to examine $R : R \in \{Results\}_\pi, R \notin \{Results\}_\partial$ *id est* results which have not previously been reviewed. Simply put, if a duplicative result is returned one would not re-review the physical documents if they have already been reviewed, but if a result is returned which has not been reviewed, obviously it then should be.

[6]Note: generally speaking, in the context of legal scholarship the Defendant is referred to with ∂ and the Plaintiff with π.
[7]Query crafted by the defense.
[8]Revised query from the plaintiff.
[9]I will mark this with emphasis, using it here as a legal "term of art" that the plaintiffs were entitled to the document as a matter of law.

The unspoken danger lurking here, from the defense's vantage point, is that more effective queries will return numerous unreviewed and unproduced results. The defense makes several inquiries as to whether I was aware or had seen various information regarding its prior DISCOVERY productions. (p 23-26) The defense's points essentially involve the large number of responsive documents already produced, and attempts to articulate *via* its cross-examination that this makes it likely new searches would not be effective. This argument is *ignoratio elenchi* because the dispute is not about the defense failing to produce relevant documents, but rather that there is relevant potential for responsive documents to be omitted due to the defects in the search queries.

The plaintiff side here likes to analogize searching to a net, and speak of the need to cast a wide net and winnow down the search from the broader starting point. I think, instead, one might suggest the design of the queries as originally crafted by the defense to be like a net with tears in it through which some of the fishes escape. The plaintiff side's point is then not that the defense produced no fishes or even an insubstantial number of fishes, but that not all the fishes to which they are entitled were caught. Further still, the plaintiff is not suggesting that every last fish must be caught, but only that a net in proper repair be used in the attempt.

The defense rounds out its point by asking if I was aware of any specific document said to exist, but which was not produced. (p 26) This is reminiscent of the Zubulake cases themselves. The court takes notice of this line of logic and interjects a question as to how it would be known or not known if further results were to be obtained unless the revised queries were run for comparison. (p 27) This is the point where the court's opinion seems to shift. If the revised searches are run, the results can be compared to the prior queries. Either new results are returned or they are not. This ease of comparison combined with the short length of time it would be necessary for the plaintiff side to craft the new queries and the defense side to run them, led the court to grant the motion. (p 27-52).

6.2.0.3 The law governing discovery

The bulk of arguments against running the revised search queries or reopening discovery are offered by the defense toward the middle of the transcripts. (p 35-43) The defense is very clear in its view that the decision of the plaintiff's side to contest the sufficiency of the search queries was opportunistic and hypocritical with the real purpose being to, "get this trial pushed off and create a big ruckus." (p 36) The defense makes several statements relating to motive, the burden of proof, and so forth in reference to Louisiana Code of Civil Procedure Article 1461. Upon reviewing the article in question, I believe the defense meant to refer to Article 1462(E) governing when a party believes a production of electronically stored information is not in compliance with what was requested. (p 36, 40, 43)

Art. 1461. Production of documents and things; entry upon land; scope

Any party may serve on any other party a request (1) to produce and permit the party making the request, or someone acting on his behalf, to inspect, copy, test, and sample any designated documents or electronically stored information, including writings, drawings, graphs, charts, photographs, phono-records, sound recordings, images, and other data or data compilations in any medium from which information can be obtained, translated, if necessary, by the respondent through detection and other devices into reasonably usable form, or except as provided in Article 1462(E), to inspect and copy, test, or sample any tangible things which constitute or contain matters within the scope of Articles 1422 through 1425 and which are in the possession, custody, or control of the party upon whom the request is served; or (2) except as provided in Article 1462(E), to permit entry upon designated land or other property in the possession or control of the party upon whom the request is served for the purpose of inspection and measuring, surveying, photographing, testing, or sampling the property or any designated object or operation thereon, within the scope of Articles 1422 through 1425.

Acts 1976, No. 574, §1; Acts 2007, No. 140, §1.

Art. 1462. Production of documents and things; entry upon land; procedure

[...]

C. A party who produces documents for inspection shall produce them as they are kept in the usual course of business or shall organize and label them to correspond with the categories of the request. If a request does not specify the form or forms for producing information, including electronically stored information, a responding party shall produce the information in a form or forms in which it is ordinarily maintained or in a form or forms that are reasonably usable. When electronically stored information is produced, the responding party shall identify the specific means for electronically accessing the information.

D. Unless otherwise ordered by the court, a party need not produce the same information, including electronically stored information, in more than one form.

E. If the requesting party considers that the production of designated electronically stored information is not in compliance with the request, the requesting party may move under Article 1469 for an order compelling discovery, and in addition to the other relief afforded by Article 1469, **upon a showing of good cause by the requesting party, the court may order the responding party to afford access under specified conditions and scope to the**

requesting party, the representative of the requesting party, or the designee of the court to the computers or other types of devices used for the electronic storage of information to inspect, copy, test, and sample the designated electronically stored information within the scope of Articles 1422 and 1425.

Acts 1976, No. 574, §1. Amended by Acts 1982, No. 451, §1; Acts 2007, No. 140, §1; Acts 2010, No. 185, §1; Acts 2010, No. 682, §1, eff. Jan. 1, 2011.

Deconstructing the rhetoric, the defense side's statements seems to follow two streams of argument. The first stream seems to be that the plaintiff side should have engaged me to construct their own search queries if they were going to engage me to critique the defense search queries. In essence, the defense seems to want to articulate the plaintiff side should at least hold themselves to the same standard they seek to hold their opponents to. I see some support to this in federal case law. (p 35-36) [35] In fact, this is very much in line of how I see case law trending and what I think optimal practice should entail. That being said, the plaintiff side produced their search queries to the defense just as did the defense to the plaintiff, but the defense either did not bother to engage their own expert or was unable to find sufficient fault with the plaintiff side's queries to raise the same issue under 1462.

The second stream seems to question where the burden of proof lies. The defense essentially asserts the requirement under 1462 for a "showing of good cause" should require the plaintiffs to demonstrate specific responsive documents not produced. It phrases the requirement to run revised searches for comparison as a sanction, objecting even before the court ruled on the matter. (p 40-41) The plaintiff side argues more persuasively that the circumstances indicating document production was not complete prompted them to seek my opinion, and my opinion of the search queries' insufficiency was a sufficient showing of good cause. (p 39-40)

The defense sums up its arguments by painting the anticipated order to run revised search queries as, "a radical new rule of law that the case law does not support." (p 40) They phrase the situation as allowing the plaintiff side to scrutinize every minutiae of the defense side's actions and require perfect accuracy. The defense evokes imagery of the "old fashioned way" discovery used to be done with lawyers searching warehouses of documents as, "they did in the old days." (p 41) I think this exchange is, at best, the last gasp of a defeated position appealing to a simpler time. Clearly there is case law to support the plaintiff's position as the circumstances here are very similar to what occurred in the Harris v BP case, and what is asked for is on par with the remedy there.[35, 27, 28] Further, as technology progresses change to legal process must come in recognition thereof. In the time the defense waxes poetic over, all this discovery and review would be conducted by hand; given how doubtful it would be the defense would like to conduct document review by

the pallet instead of by electronic review tool, I would find it curious their preferred method of conducting discovery would be galavanting about dusty warehouses instead of searching electronic indices.

6.2.1 Conclusions

The danger of lax search practices producing indefensible results is very real. Here we saw the problem of a deficient search laying dormant and undetected until the eve of trial only to manifest itself like the opening of Pandora's box. The hearing here was on 11/03/2010. The revised queries were prepared and sent to the defense on 11/04/2010, and they responded with the results on 11/08/2010. At the time of this writing, the trial has still not transpired. The defense noted the discovery for this case occurred over the course of six years; that those six years could be undone so easily by the application of the same boolean logic taught to undergraduate students is shocking in the least. In light of the dire consequences at stake, the legal field would do well to adopt the security rallying cry, "Trust, but verify."

Chapter 7

Case Study: Henry Properties v Apache Corporation

7.1 Henry Properties v Apache Corporation[1][2]

7.1.1 Description

7.1.2 Exhibit A / Letter to Plaintiffs' Counsel

The letter from me to the Plaintiffs' counsel which was transmitted *via* email to the BP's counsel and originally attached to my affidavit (7.1.3) as Exhibit A. The letter is fairly straight forward informing Plaintiffs' Counsel of a discrepancy I found between what the Defense represented their search queries were run as, and what results the represented queries should have returned. In particular, the issue here, aside from normal disputes over search breadth, involved a proximity connector used in the Concordance[3] system. The "near" connector is specified as $nearX$ where $X : X \in \mathbb{Z}^+$ and will match the term before and after it if they occur within[4] X terms of each other.

In the instance here, the Defense gave the Plaintiffs search result records from their searches after a meeting we had at the Defense Counsel's office. In the meeting, I asked for clarification as to what connector

[1] Apache Corporation is the predecessor in interest to BP

[2] Full case heading is C. F. HENRY PROPERTIES, LLC, SHERYL LEBLEU SWIFT, MARSHA LEBLEU DELANEY, JOHN HENRY LEBLEU, II, LOREE WARE LEBLEU, ARMANT MARK LEBLEU, SCOTT HENRY LEBLEU, PAIGE LEBLEU, INDIVIDUALLY AND ON BEHALF OF THE MINOR, CAMERON WAYNE LEBLEU, GLENN CONWAY LEBLEU ESTATE, WILLIAM F. HENRY, JR. ESTATE, EDWIN SCOTT HENRY, SARAH ANN HENRY, MARY ELLEN HENRY, JAMES ADAM HENRY, JOHN DAVIS HENRY, CHARLES GREGORY HENRY, CANDACE HENRY OLIVER, HUTCH P. HENRY, PETER C. HENRY, JR., JANE ANN HENRY, JANE ANN HENRY TRUST, PATRICIA HENRY, PATRICIA HENRY TRUST, BENNIE JANE D. HENRY ESTATE, MATTHEW CHARLES LOPEZ, DAVID ALLEN LOPEZ, REBECCA LYNN LOPEZ RABB, MARK ALEX LOPEZ, CORINE A. LOPEZ McGILBERY, SALLY LOPEZ WELCH, BETTY JEAN HENRY, CAROL ANN MARTIN, EDWARD CAYO MARTIN, HILDA P. HENRY, ELLRAY JAMES HENRY, MELANIE HENRY HEBERT, MICHAEL S. HENRY, CLAY B. ADAMS, KAREN ADAMS HARRIS, THOMAS E. ADAMS, GEORGIA CORBELLO, JEFFREY A. CORBELLO, KIMBERLEE CORBELLO PALERMO, MADELYN SUE STACY MCCALLISTER, JERRY LYNN STACY, ROBERT C. CORBELLO, RICHARD A. CORBELLO, BENNIE W. CORBELLO, PAULA LAROQUE LAWLIS, SUSAN LAROCQUE JACKSON, THOMAS ARTHUR LAROCQUE, HARRY JAY LAROCQUE, MONTE M. HURLEY, DOROTHY WHITE ROSTEET ESTATE, JOSEPH WILBUR ROSTEET, JR., MICHAEL J. ROSTEET, AND MATTHEW J. ROSTEET VS. APACHE CORPORATION

[3] Concordance is a database / search system owned by Lexis Nexis. BP uses it as their search index for physical documents.

[4] Order does not matter. So A near2 B will match A before B, and A after B so long as A and B are within 2 terms of each other.

their $\backslash x$ notation was representing; they clarified it was being "translated" by the person running the searches from the Counsel's short hand notation to Concordance syntax as $nearX$. (Affidavit, p 3) I loaded the data produced by the Defense's Counsel into a Concordance database to verify their search results. Discussion continues at (7.1.3).

June 3, 2011

Brian Roux
Venue Docket, LLC
701 Poydras St, Suite 150P
New Orleans, LA 70139

[redacted]

Dear [redacted]:

The letter from Mr. [redacted] dated May 31, 2011 clarifying the proximity search connectors indicated all searches identified as "John w/3 Doe" or as "John /3 Doe" were run by BP as "John near3 Doe". Based on this information, I have examined the spreadsheet result sets we were provided.

In examining the file, **Hard Copy Box Log_Ardoin_Version 2_12022010_1.XLS**, I found a discrepancy between the reported number of hits for the **Cameron /2 Field** query and the number of hits returned under the query BP would have run under the above representation (**Cameron near2 Field**). The query as reported had 0 hits. After importing the spreadsheet data into a Concordance database, I found just the records in the indicated spreadsheet contained 40 hits for **Cameron near2 Field** (where order does not matter), and 10 hits for **Cameron adj2 Field** (where order does matter). Running the search literally as written (**Cameron /2 Field**) was the only form of this search which produced a 0 hit count. Let me reiterate, my test searches encompass only the records which were present in the **Hard Copy Box Log_Ardoin_Version 2_12022010_1.XLS** file; the hit count for **Cameron near2 Field** run properly in the database may be substantially larger.

For example, if you refer to the entry for Box ID IBM00075810, the title field contains the following information (emphasis added): *CODE 390 SETTLEMENT BOARDS **CAMERON FIELD**, EAST HOLLY BEACH, JENNINGS TOWNSITE, SOUTH TOWNWELL, YSCLOSKEY PLANT, LAKE RACCOURCI, SHIP SHOAL-177, WEST DELTA-35, CAILLOU ISLAND, BASTIAN BAY (11-1988 TO 06-1991)*

Based on this discrepancy and my ability to reproduce these results in Concordance, the same system used by BP for its searches, I must conclude the searches were run incorrectly or the proximity connectors were not properly translated into Concordance's 'near' notation. All searches which use invalid notation should be rerun using proper query syntax.

If you have any questions, please do not hesitate to contact me.

Warm regards,

[redacted]

Brian Roux [redacted]

7.1.3 Affidavit

When I translated the short hand search notation into proper Concordance syntax and ran the searches the Defense represented they used I found that searches they claimed zero hits for actually produced numerous hits within just the dataset they gave us. After attempting other possible versions of the translated search which were order dependent, I found the only rationally conceivable representation which was consistent with the results they presented to us was to run the searches literally in the short hand notation without translation. (p 4)

In total, just the data provided to us as the results of other searches, there were 932 instances which would have been flagged by properly translated search queries. (p 4) This brought into question every single search performed in the case thus far, and potentially indicated problems in many other cases depending where in the chain the failure occurred. (p 5)

C. F. HENRY PROPERTIES, LLC, SHERYL LEBLEU SWIFT, MARSHA LEBLEU DELANEY, JOHN HENRY LEBLEU, II, LOREE WARE LEBLEU, ARMANT MARK LEBLEU, SCOTT HENRY LEBLEU, PAIGE LEBLEU, INDIVIDUALLY AND ON BEHALF OF THE MINOR, CAMERON WAYNE LEBLEU, GLENN CONWAY LEBLEU ESTATE, WILLIAM F. HENRY, JR. ESTATE, EDWIN SCOTT HENRY, SARAH ANN HENRY, MARY ELLEN HENRY, JAMES ADAM HENRY, JOHN DAVIS HENRY, CHARLES GREGORY HENRY, CANDACE HENRY OLIVIER, HUTCH P. HENRY, PETER C. HENRY, JR., JANE ANN HENRY, JANE ANN HENRY TRUST, PATRICIA HENRY, PATRICIA HENRY TRUST, BENNIE JANE D. HENRY ESTATE, MATTHEW CHARLES LOPEZ, DAVID ALLEN LOPEZ, REBECCA LYNN LOPEZ RABB, MARK ALEX LOPEZ, CORINE A. LOPEZ McGILBERY, SALLY LOPEZ WELCH, BETTY JEAN HENRY, CAROL ANN MARTIN, EDWARD CAYO MARTIN, HILDA P. HENRY, ELLRAY JAMES HENRY, MELANIE HENRY HEBERT, MICHAEL S. HENRY, CLAY B. ADAMS, KAREN ADAMS HARRIS, THOMAS E. ADAMS, GEORGIA CORBELLO, JEFFREY A. CORBELLO, KIMBERLEE CORBELLO PALERMO, MADELYN SUE STACY MCCALLISTER, JERRY LYNN STACY, ROBERT C. CORBELLO, RICHARD A. CORBELLO, BENNIE W. CORBELLO, PAULA LAROQUE LAWLIS, SUSAN LAROCQUE JACKSON, THOMAS ARTHUR LAROCQUE, HARRY JAY LAROCQUE, MONTE M. HURLEY, DOROTHY WHITE ROSTEET ESTATE, JOSEPH WILBUR ROSTEET, JR., MICHAEL J. ROSTEET, AND MATTHEW J. ROSTEET	:	38th JUDICIAL DISTRICT COURT
	:	STATE OF LOUISIANA
VS.	:	PARISH OF CAMERON
APACHE CORPORATION	:	CASE NO. 10-18683
Filed: _____	:	_____ Deputy Clerk

AFFIDAVIT OF BRIAN ROUX

BE IT KNOWN that before me, the undersigned notary, duly commissioned and qualified in the State of Louisiana, came and appeared BRIAN ROUX, who, after being sworn, did depose and say as follows:

1.

I am over 18 years of age and under no legal disability. The facts and opinions stated in this Affidavit are true and correct, and are based upon my own personal knowledge.

2.

I am an information technology specialist, with extensive experience in database architecture and operations, as well as forensic investigation of information technology ("IT") systems. I am currently employed by Digital Inquest, LLC, but until several weeks ago, I was employed by Venue Docket—a firm specializing in electronic discovery and located in New Orleans, Louisiana.

3.

I routinely provide consulting services to attorneys on electronic discovery matters. In fact, I have previously provided consulting services in several cases involving the defendant, BP America, Inc. ("BP"), and thus I have gained extensive knowledge regarding BP's IT systems and, in particular, its electronic file database.

4.

BP's electronic file database allows textual searches of only the descriptive file index records, as opposed to textual searches of the documents themselves. These descriptions are prepared by BP's employees or agents, who enter some—but apparently not all—of the archived files into the database.

5.

In other words, BP's electronic file database contains at most a limited description of each archived file or folder, but not a description of each document contained within that file. Again, BP's database only searches these descriptions of the files or folders, which often consist of only a few words.

6.

For this reason alone, and based on my prior experiences with BP and its electronic discovery efforts and team, BP's electronic file database is not a reliable means of searching for and locating documents in response to discovery requests served in litigation.

7.

Putting aside the unreliability of BP's electronic file database, the electronic searches BP has conducted in this case on its database are ineffective and deficient.

8.

On May 31, 2011, I attended a meeting between counsel for the plaintiffs and BP at the offices of BP's counsel, Kean Miller. The sole purpose of this meeting was to discuss the discovery issues in this case. At that meeting, BP provided the plaintiffs with electronic copies of several spreadsheets purportedly showing the files that were triggered as a result of BP's search terms.

9.

After reviewing these spreadsheets, which I only later learned contained missing or omitted columns of information, it became clear to me that the searches BP had conducted were deficient for several reasons.

10.

First, the search queries BP used to locate potentially responsive documents were ineffectively tailored for the generic information input into BP's electronic file index. For example, BP reported that it received no hits when it searched "(ground w/2 water) AND (contamin*) AND (Louisiana or La)." However, the plaintiffs' modified searches—which were more appropriately designed to locate all potentially responsive material—for "(groundwater) AND (Louisiana OR LA OR Cameron)" and "(ground near3 water) AND (Louisiana OR LA OR Cameron)" returned a total of 557 hits.

11.

As another example, BP reported that it received no hits when it searched "James Henry Company." However, the plaintiffs' modified search for "(James OR Company) near3 Henry" returned a total of 139 hits. BP also reported that its search of "Cameron SWD" resulted in 7 hits, but the plaintiffs' modified search for "Cameron near5 SW*" returned a total of 85 hits.

12.

It bears noting that the examples set forth above are merely illustrative. There are numerous other examples that show the deficiencies associated with the search terms BP ran on its electronic file database.

13.

Second, after reviewing the spreadsheets showing the purported results of BP's searches, it became clear to me that all of BP's proximity searches were run not correctly run after I was unable to reproduce BP's search results in Concordance, which is the database BP uses to run its search terms. Specifically, BP's search reported that it received no hits when it searched "Cameron /2 Field," even though one entry for BOX ID IBM00075810 contained the following description: "CODE 390 SETTLEMENT BOARDS **CAMERON FIELD** . . . ." (Emphasis added). If BP's search related to Cameron Field had been properly run, it obviously would have located this file entry.

14.

I set forth in detail these discrepancies in a letter to the plaintiffs' counsel dated June 3, 2011, in which I requested that all of BP's searches using proximity connectors should be rerun using the proper query syntax. A copy of the e-mail from the plaintiffs' counsel to BP's counsel transmitting my June 3 letter is attached hereto as **Exhibit A**.

15.

Pursuant to my request, BP re-ran all of its searches containing proximity connectors. A table depicting the number of hits based on BP's initial searches and its revised searches using the proper query syntax is set forth below.

Initial Search Term	Initial Hits	Revised Search Term	Revised Hits
Cameron /2 Field	0	Cameron near3 Field	494
Sarah /3 Henry	0	Sarah near3 Henry	3
Edwin /3 Scott	0	Edwin near3 Scott	3
Mary /3 Henry	0	Mary near3 Henry	13
John /3 Henry	0	John near3 Henry	136
Charles /3 Henry	0	Charles near3 Henry	47
James /3 Henry	0	James near3 Henry	99
Patricia /3 Henry	0	Patricia near3 Henry	2
Betty /3 Henry	0	Betty near3 Henry	9
Carol /3 Martin	0	Carol near3 Martin	8
Edward /3 Martin	0	Edward near3 Martin	37
Michael /3 Henry	0	Michael near3 Henry	28
Clay /3 Adams	0	Clay near3 Adams	3
Thomas /3 Adams	0	Thomas near3 Adams	14
Susan /3 Jackson	0	Susan near3 Jackson	5
Jane /3 Henry	0	Jane near3 Henry	8
Matthew /3 Lopez	0	Matthew near3 Lopez	1
David /3 Lopez	0	David near3 Lopez	12
Mark /3 Lopez	0	Mark near3 Lopez	1
Karen /3 Harris	0	Karen near3 Harris	9
TOTAL HITS	0	TOTAL HITS	932

16.

As shown above, the ineffectiveness of the searches performed by BP is evident, whether such deficiencies exist by design or are the result of inadequately trained IT technicians who have only a rudimentary knowledge of BP's electronic file database, its search mechanisms, operations, and syntax. However, this is not the first case in which I have uncovered deficiencies related to the electronic searches performed by BP for purposes of locating documents during the course of discovery.

17.

These repeated problems with BP's electronic discovery efforts call into question the reliability and effectiveness of every search conducted by BP on its electronic file database for purposes of locating documents in response to the plaintiffs' discovery requests.

I declare under the penalty of perjury that the foregoing statements are true to the best of my knowledge and belief.

_____

BRIAN ROUX

SWORN TO AND SUBSCRIBED before me, Notary Public, on this _____ day of July, 2011.

_____

NOTARY PUBLIC

7.1.4 Henry Properties v Apache Corporation Conclusions

Henry v Apache represents another hidden danger in the developing areas of electronic discovery - familiarity with the specifics of the underlying technologies used in the search system. The problem found here of the Defense failing to translate the short hand notation into a valid search query shares pitfalls of the earlier example involving saltwater. To the human eye, a consistent notation which is symbolically plausible can easily escape notice even though erroneous. This problem is compounded when we consider the $\backslash x$ notation may be valid in other systems; in fact, it is valid in Westlaw, the legal search engine commonly used by attorneys. If we assume this is the reason the error was not caught - familiarity with search systems for which it is valid, and naive assumption this syntax might be universal - it still leaves the troubling fact this was not corrected even after clarification was requested regarding the notation's meaning and use in the Concordance searches.

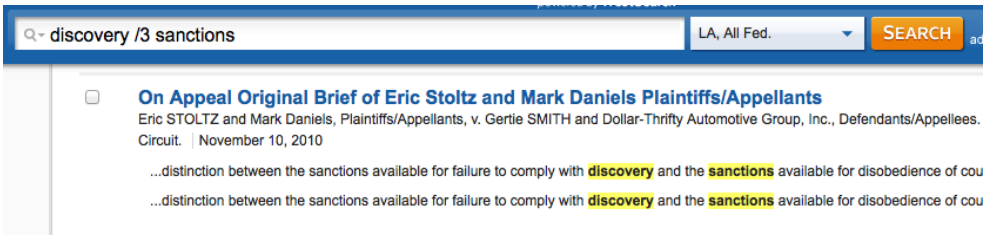


Figure 7.1.1: Screen capture of the $\backslash x$ notation working in Westlaw.

I will now further refine the functional representation of an act of searching I introduced *supra*, to extend the functional representation to include the normal practice of conducting numerous searches in discovery. (Affidavit, p 4) Recall the representation $S_j(Q_i, A_j(U, \{Corpus\})) = \{Result\}$. This represents one search query Q_1 which in the case here may have been "cameron near3 field". There were twenty searches using invalid notation cited in my affidavit here. (Affidavit, p 4) We might term these $\{Q_1, Q_2, Q_3 \ldots Q_{19}\}$ or more generally $\{Q_1, Q_2, Q_3 \ldots Q_n\}$. It would make sense, then, to represent the searches for these revised queries as:

$$\sum_{i=1}^{n} S_j(Q_i, A_j(U, \{Corpus\})) = \{Result\}$$

Figure 7.1.2: Functional representation of a series of searches in DISCOVERY

Considering now the problem with syntax translation, we might assume that the individual specifying the search had a boolean query in mind but lacked knowledge of the search system (or was unaware the queries were expressed in the wrong system's syntax). The functional representation may then be further refined by introducing a function T_j which will translate any query Q_i such that its logical representation is valid syntactically for system S_j. Thus, the functional representation becomes:

$$\sum_{i=1}^{n} S_j(T_j(Q_i), A_j(U, \{Corpus\})) = \{Result\}$$

Figure 7.1.3: Revised functional representation of a series of searches in DISCOVERY.

As the DISCOVERY process is reduced to a functional representation, we see the complexity more clearly. The preceding case studies deal with conflicts over a single and specific system S_j because in the circumstances of those instances, that was the only system involved in the conflict. In the DISCOVERY process there is more likely to be a larger heterogenous set of systems needing to be searched. As we enumerated the queries so to should we enumerate the search systems $\{S_1, S_2, S_3...S_m\}$. This leads naturally to a final refinement of the functional representation:

$$\sum_{j=1}^{m} \sum_{i=1}^{n} S_j(T_j(Q_i), A_j(U, \{Corpus\}_j)) = \{Result\}$$

Figure 7.1.4: Functional representation of a series of searches in DISCOVERY conducted over a series of searchable systems.

Now the complexity continues to increase. We can describe the complexity in terms of $\#\{S\}$, $\#\{Q\}$, and $\#\{Corpus\}$, generically resulting in $T(n) = O(n^3)$ as a base.[5] If $\#\{S\}$ is reduced to 1 then $T(n) = O(n^2)$. The only unknown complexity is, *ergo*, the individual complexity of a Q_i run in S_j; where such complexity exceeds that of the system or worse pushes it to another complexity class, it might be termed either not feasible or in need of revision. This is often the case with Q_i which use wild card prefix or postfix notation[6].

The case studies *supra* examine disputes where there is a single, pre-existing document *corpus* and search system. This is not generally the case, and was not the entire case in Harris v BP. Not included in that case study was a dispute over email databases of which there was potentially both a legacy system and a modern system. In general incidence, a DISCOVERY will encompass document database systems, email systems, documents on individual computer systems and the like. Some of these data systems may have S_j coupled with the $\{Corpus\}_j$ but others will just be an unbound $\{Corpus\}$ lacking S to search with. As I will introduce *infra*, there are numerous systems available for review, but none in the Open Source tool chain. I seek to cure such in Part III.

[5]Each $\{Corpus\}_j$ is specific to the S_j otherwise running every $\{Corpus\}_k$ through S_j would increase $T(n) = O(n^4)$ and be highly duplicative.

[6]In my experience, lawyers often call this root expansion in the later case. I dislike the root expansion terminology because in some systems the root expansion and a postfix wildcard may be identical, but in others root expansion may rely on some linguistic sense of root expansion *exempli gratia* Base* expanding to bases and baseball, but not to a non-dictionary term BASE3114 or some such. In cases where root expnsion has such a meaning, there may be a real difference between root expansion and wild card.

Case Study Conclusions

The three case studies presented herein provide a novel view of problems arising in discovery, the causes of those problems, and a view over time showing the problems are not being addressed. The Harris case involves two major oil companies as defendants; these companies have vast resources at their disposal, and sophisticated legal teams regularly handling litigation. At both BP and Conoco there is a clear deficiency in the institutional knowledge possessed by the respective defense legal teams. The problems identified in the Livelink system returning different result set sizes on identical queries run under the same user, and dramatically larger result set sizes when those queries are rerun with super user credentials show that not only is a sufficient search not being conducted, but also that the defense is mistaken in their assumptions to the contrary.

The VPSB case shows similar concerns about user access and search query sufficiency, but with Chevron. In both the opposition brief and at oral argument, the defendant's counsel argue not about why they think their searches are sufficient, but instead argue the quantity of documents produced should indicate sufficient searches. However, as demonstrated by the events in Harris, the quantity of documents is not relevant if poor search practices or insufficient user access are present. A bad search is, by analogy, like leaving rooms investigated because other rooms have already been found.

Apache brings the case studies full circle back to BP again. Again, concerns about search query sufficiency are raised regarding the discovery searches. In addition to inadequately designed searches, I discovered invalid syntax was being used thus invalidating the majority of search results the defense claimed to have run. In other words, even if the searches were adequate in their scope, the results would still be wrong because the searches were translated into invalid syntax.

The flawed practices I detailed in the three case studies went undetected by numerous sophisticated parties (IT personnel, in-house counsel, outside counsel, and litigation support personnel) in three very large, well funded organizations. The problems were not cured as we see from BP's repeat failure in Apache after been put on notice by the events in Harris. This leads to the conclusion that the problems are real, are pervasive, and are not being addressed. Because the discovery process leaves the opposing side with limited knowledge of how searches are conducted, significant technical knowledge is needed to detect these

inadequacies, and it is clear that level of knowledge is not being regularly used.

Part III

New Tools

Chapter 8

Black Friar

8.1 Background of the Field

At present the field has three main phases of practice: acquisition, analysis, reporting. Acquisition originated in dead acquisition where the data storage medium, such as a hard drive, is imaged byte-for-byte to produce an exact duplicate when the system is powered off. The duplicate is hashed for later verification after analysis is complete. In a more modern twist, Live analysis involves acquiring data from a system while it is still running. Live acquisition allows for preserving more ephemeral data such as memory dumps, active network connections, logged on users, running programs, etc which would otherwise be lost in powering the system down for dead acquisition. Live acquisition risks the triggering of anti-forensics tools, malicious commands from still logged in users, and damaging the system state.

However the storage data is acquired, the image files are transferred to a tool suite where the examiner/analyst/investigator/researcher etc begins the careful analysis process. Depending on the size of the datasets, analysis may take a very long time to complete. Whether the examiner uses a tool suite, or a collection of individual tools some common tasks will be carried out. Files will be hashed to exclude known good files (operating system libraries, known executables, etc), suspicious files will be flagged (erroneous file extensions, unexpectedly large files, encrypted or password protected files) for closer scrutiny, and text will be searched for relevant keywords. When the examiner finishes the analysis, a report is drafted summarizing and interpreting the evidence ready for consumption by lawyers, law enforcement, and the courts.

8.1.1 What problems do we face?

The most prevalent problem we face is continuing increases in dataset sizes without equivalent increases in transfer bandwidth. In the past few years, the capacity of consumer hard drives has increased 400%, and there are more sources of data than ever before (see figure 8.1.1). Where hard drive storage was once scarce

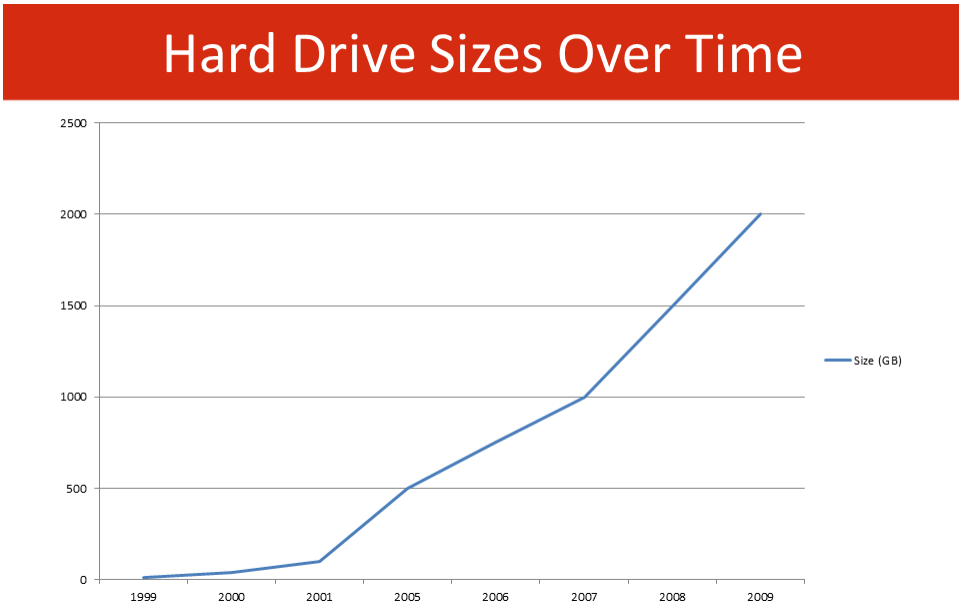


Figure 8.1.1: Hard Drive Size Increases

and expensive, it is not cheap and plentiful. In both consumer and business spaces the number of data sources has increased to include external hard drives, thumb drives, cell phones, cameras, gaming systems, printer/copiers, and Network Attached Storage. Bandwidth, on the other hand, has never enjoyed more than a 100% increase in bandwidth between generations (see figure 8.1.2) and, in some cases, has had meager gains $< 20\%$ (IDE133 to SATAI). Even when bandwidth increases occur, there are generally still differences between theoretical maximum bandwidth and average utilization for a given transfer.

Increasing target sizes would not be prohibitive if the other required resources needed for an investigation increased proportionally. Putting aside access bandwidth concerns, the primary bottleneck in investigations which cannot be easily remedied is the human component. Bandwidth can be artificially increased by spreading the load over multiple drives, caching techniques, and other technological solutions, but the human factor cannot be quickened. The graduation rate of Computer Scientists has fallen at all levels (B.S., M.S., Ph.D.) since 2004 (see figure 8.1.3), and even of those graduated the number focused on digital forensics is a small fraction.

In a nutshell, we are facing increasing dataset sizes, access speeds which do not increase at the same rate as capacity increases, and a drop in production of humans with sufficient training to perform analysis. All these factors increase the turnaround time from the start to the conclusion of an investigation and the production of its findings to the next stage.

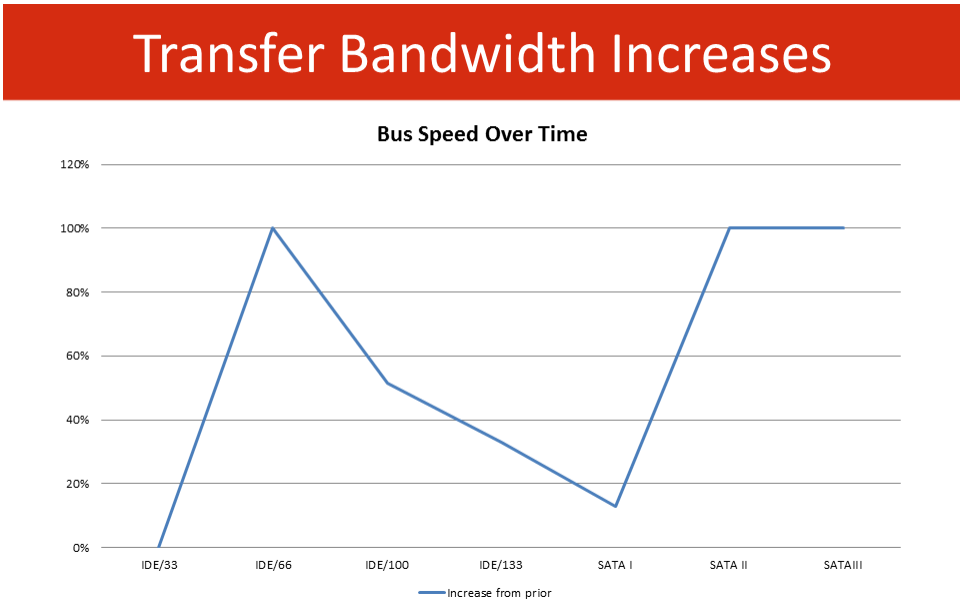


Figure 8.1.2: Transfer Bandwidth Increases Over Prior Generations

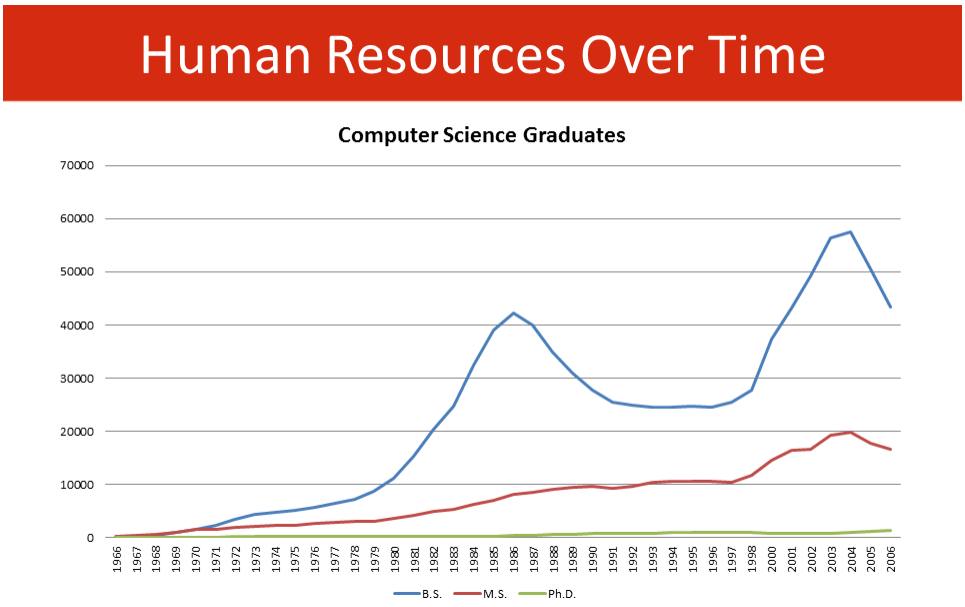


Figure 8.1.3: Human Resources

8.1.2 What are our tools?

Many of the tools we use started out intended for system administration. More specific tools were created later such as foremost/scalpel for file carving, and stegdetect for stego detection. Many tools are created to handle a specific task or situation encountered in the course of an investigation. The majority of these tools were designed with a single system / single thread in mind. As the tool sets matured, tool suites became available which combined single use tools into user friendly packages with GUIs to guide the investigator through the analysis process and carry out rudimentary case management.

Both sets of tools fall prey to scaling issues. They focus on a single investigator, do not provide for collaboration between investigators, and are highly dependent on the investigator to report findings to interested parties down the chain. They also suffer from updating issues in that coupling an individual tool into a suite limits the ability for the investigator to run the most updated tool between suite releases. Ayers in the 2009 proceedings of the Digital Forensics Research Workshop (DFRWS) identified a set of features he felt required to consider a forensics tool to be "second generation." I differ slightly in my requirements by giving greatest precedence to distributed/parallel processing, collaboration, application specific extensibility, and increased focus on a pipeline approach to decrease turnaround time.

8.1.3 Where do we fit in?

We must recognize our role in examining a dataset is one part of a greater process. The end of the process lies with the law enforcement, attorneys, and ultimately the courts. Those parties cannot begin their work until they are furnished with our work, but they do not necessarily require all of our work to begin theirs. We can view our current process as a synchronous one in which everyone waits for us to complete, but by moving to a more asynchronous pipeline our partial results can be fed into the pipe so law enforcement and attorneys can begin building their cases in the courts. Ultimately those interested parties will require our final results, but a significant amount of preliminary work can be completed before the final report is submitted. Additionally, later parts of the pipeline can provide feedback to the investigator for better or more precisely targeted investigations.

8.2 Black Friar

Black Friar is an experimental prototype of the distributed/parallel pipelined process. Its aim is to distribute the initial processing, front load as many computationally expensive tasks into the first file read as possible, provide an architecture capable of being extended to handle processing specific to proprietary file types

in either the first pass or in a more exhaustive second pass, and allow for usable intermediate results to be shared among investigators for collaborative purposes or fed into the next stage of the pipeline. While proprietary suites are starting to use distributed processing, this capability is lacking in open source tools. The prototype, therefore, seeks to leverage the existing tools in such a way as to make tasks which use them distributable while retaining the ability to upgrade the individual tools to the latest version independently of the framework.

The current prototype used for these results extracts data from drive images using The Sleuth Kit. It distributes the file load over the nodes producing file hashes, extracting string data, and identifying file types. The document information along with the extracted string data is formed into an index using the Lucene project. The Lucene index produced by each node is fully usable on its own prior to merging allowing segments of interest to be immediately used before the overall process is complete.

The test systems included a Xeon quad core work station with 8gb of ram and I/O spread over multiple drives, and a 7 node cluster of older dual Xeon dual core based systems with 2 gb of ram and single drives. Results show the performance difference between the quad core system and an individual node at each thread number. In both cases performance gains were significant from 1 to 2, and 2 to 3 threads but not between 3 to 4 where the process again became I/O bound. Experimental results are separated into processing which generates the hashes, determines file type, extracts string data, and queries file metadata from the image, and indexing which combines the processing results into a Lucene index. In the reported processing results, the distributed metrics include the image file being present on each individual node and the image file being stored on the Gluster File System as communal storage with the systems interconnected with gigabit ethernet. Performance decreases using gluster as opposed to storing a copy locally on each node was minimal and presumably would disappear with an interconnect with more bandwidth.

On the processing side, the fully distributed processing was 18% of the single thread single node processing time and 26% of the time when compared against a fully threaded single node run. Indexing on the cluster was 9% of the time required for a single threaded single node run, and 14% of a fully threaded single node run. These results are derived from the DC3 2009 data set as a target.

As a prototype, Black Friar was designed to handle hard drive images in an investigation context. As exemplified in the Part II case studies, much of my professional work revolves around electronic discovery where the current standard output is the paginated tiff or pdf rendering of native format files. In the context of a large data collection from a corporate setting where no central indexed file storage exists, Black Friar would be useful for triaging a discovery effort. A large collection of files could be centralized for preservation, and pertinent keyword searches designed to identify responsive discoverable material could be run against the index without the wasted effort of converting and manually reviewing rendered versions of all native

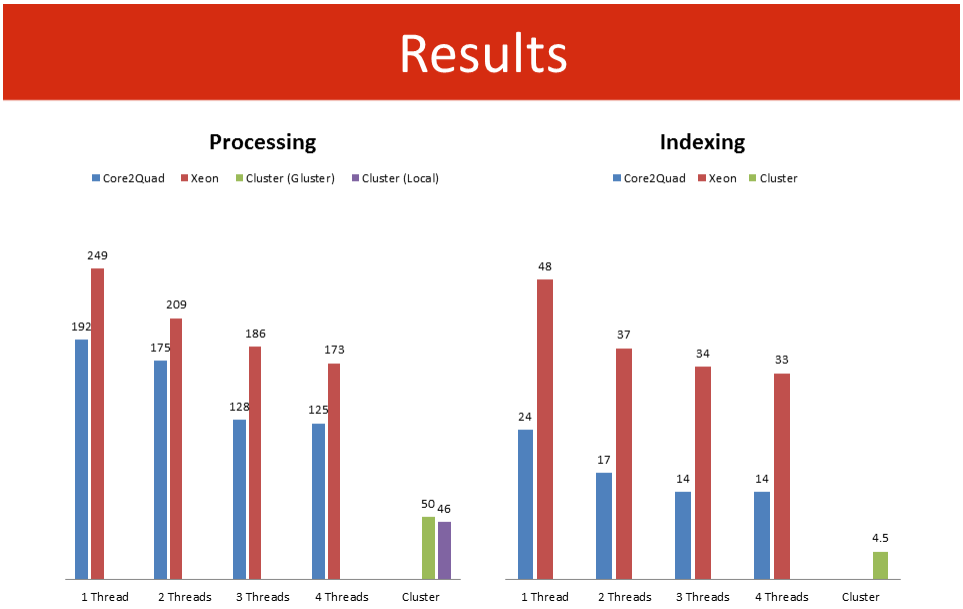


Figure 8.2.1: Black Friar Results (Time in Min)

electronic files.

In keeping with a pipelined workflow, files identified in discovery searches can be exported for conversion, attorney review, and production. Such productions could be updated simply by rerunning the queries as more files are collected, and continuing to feed new results to the conversion and review workflow. Finally, in the event of discovery disputes the query results can be reproduced, can be augmented with additional searches, and search sufficiency can be defended by explaining the search design in the context of a publicly documented query language.

Black Friar's weakness is being designed for distributed processing using a Linux cluster with a distributed file system. This construction makes Black Friar a powerful tool for triaging and investigating large data sets, but too large of a tool for small to medium sized data sets or for organizations without the internal technical resources or personnel to administrate such a system. Using what I learned from my experiments with Black Friar, I decided to examine the existing proprietary solutions for converting electronic files to tiff and pdf paginated formats and the subsequent review tools used to manually identify responsive documents; the aim of this examination was to develop an equivalent tool suite which could be fed by or integrated into black friar. The results of this examination are presented in the remaining chapters.

Chapter 9

ESI Processing and Discere

In Part I, I discussed the brave new world of electronic discovery. In Part II, I demonstrated *via* case studies the kinds of difficulties that exist in conducting electronic discovery and the pitfalls for the unwary. The prior chapter dealt with problems of scale, the benefits of adopting distributed approaches, and the necessity of using a feedback process to refine the search for pertinent data. In this chapter I will begin to present Discere, my tool designed to encompass the entire ESI processing workflow.

Before discussing Discere, I must define what the ESI workflow actually is, how it is currently implemented, and what flaws presently exist in existing approaches. In so doing, I will set a backdrop to compare how my approach differs and to explain why my approach is superior. I will discuss necessary changes to current industry practices for document numbering in (11.1) and (11.2) to facilitate multi-threaded and distributed processing, and then I will address distributed processing in (11.3) and (11.4).

9.1 What is ESI Processing?

Recall from *Victor Stanley*[1] that counsel is responsible for reviewing documents not only to comply with its discovery obligations by producing responsive documents to opposing counsel, but also for reviewing those documents prior to production to ensure privileged documents are not inadvertently disclosed. How does one accomplish this duty? In the physical document analogue, one would review each document prior to production by flipping through its pages. In a modern context, these documents are often scanned to allow review by computer and then produced as images or re-printed sans the excluded materials.

When ESI began appearing in discovery, lawyers tended to retreat to what they were comfortable with. To deal with electronic documents, spreadsheets, emails, drawings, and other ESI, the legal community adopted a pagination process. An email, for example, does not have pages, but is rather a stream of text[2].

[1] [35] See 3.2 on page 39 (where documents were inadvertently disclosed and the need for rigorous search practices was highlighted.)

[2] See figure 9.1.1 on the next page

Figure 9.1.1: Illustration of email pagination

If an email is printed, its content is rendered by the mail client into a paginated form appropriate to the size paper it will be printed on. Even documents intended for printing such files produced by word processors do not have an actual pagination, but rather a currently rendered pagination for the page size the document is set to. By adopting a pagination process to match the physical analogues, standard practice in the legal community became, essentially, to convert all electronic information into a discrete and final pagination prior to review, and then to review the paginated documents rather than the original electronic information.

There are multiple problems with paginating an electronic document. First, pagination is entirely arbitrary. Depending on paper size, margins, spacing, font, et cetera, the number of pages the document will have once rendered will vary. Later on, we will see this variance causes unexpected difficulties with distributed processing due to another legacy legal practice called a BATES NUMBER; I cure these difficulties by modifying the practice in question.

Second, a rendered document will potentially lose information from METADATA[3] and other non-visible data sources. This can be a significant problem if that information is relevant discoverable information and is not produced. If the native electronic format is produced, on the other hand, this information might also

[3]Metadata is an often used but seldom understood term in the legal sphere. Its meaning varies and sometimes encompasses data which should not be considered metadata at all.

Product	1-Year Pricing	2-Year Pricing	3-Year Pricing
EDD Platinum Bundle	$12,705	$10,475	$9,349
ES Platinum	$11,253	$9,275	$8,280
Scan Platinum	$9,620	$7,932	$7,079
EDD Premium	$8,168	$6,734	$6,010
Scan Premium	$6,534	$5,391	$4,808
Core Licenses Bundle	$4,356	$3,594	$3,207

Table 9.2.1: LAW PreDiscovery Bundle Pricing Structure

constitute a privilege problem if a given native format contains prior revisions which are not immediately visible, but remain accessible. This raises the question as to whether the rendered format should be considered authoritative.

Third, with the increasing quantity of ESI being produced, pagination must be automated to be feasible. Automation presents the difficulty of multiple and often proprietary file formats. Further, some file formats do not lend themselves to pagination. A database, for example, contains data which might be discoverable and relevant, but would essentially be meaningless if produced as a rendering of the tables. Such data would only have human cognizable meaning when rendered through a query, a report, or some other process which produces output for human consumption. We will see that the automated rendering process creates a significant time constraint with inferior implementations.

9.2 What are the existing approaches?

There are a number of commercial solutions, applications, and vendors who will, for a high price, render ESI into a paginated format. The current most commonly used solution is an application called Law Pre-Discovery ("LAW"). As this represents, essentially, an industry standard I will use it for comparison and analysis of problems inherent in existing approaches. It is important to understand the costs associated with these platforms, their performance and scalability, and the potential impact it has on smaller law firms or solo practitioners.

The pricing structure for LAW is represented in table 9.2.1 and table 9.2.2 on the next page. The pricing structure is an annual subscription with discounted pricing for multi-year licensing agreements. For the purposes of considering conversion of ESI from native electronic format to a paginated format, LAW requires the Admin[4], ED Loader[5], and TIFF Conversion modules[6]. This minimum functionality for a single station will cost between $4,013-5,457 annually depending on license agreement length.

The three modules required to perform the base minimum conversion to paginated format leave large

[4]Allows for base functions in creating and managing case files.
[5]Allows electronic information to be imported into a case.
[6]Allows batch conversion of electronic files to TIFF format.

Module	1-Year Pricing	2-Year Pricing	3-Year Pricing
Scan-Unlimited	$3,086	$2,547	$2,371
LAWtsi Scan	$3,086	$2,547	$2,371
ED Loader	$3,086	$2,547	$2,371
OCR (ABBYY Fine Reader)	$2,008	$1,646	$1,476
QC/Edit	$1,367	$1,125	$1,004
Tiffing	$1,367	$1,125	$1,004
Endorse	$1,004	$823	$738
OCR	$1,004	$823	$738
E-Print	$1,004	$823	$738
Print	$1,004	$823	$738
Full-Text Indexing	$1,004	$823	$738
Admin	$1,004	$823	$738
Searchable PDF	$1,004	$823	$738

Table 9.2.2: LAW PreDiscovery Module Pricing Structure

functionality gaps including an inability to print, bates number, perform quality control / edit functions, or OCR the resulting TIFF images. A more realistic option is the EDD Premium bundle which includes the necessary modules to perform these additional function. The cost ranges from $6,010-8,168 depending on license duration. This is, again, for a single work station. After obtaining the base package, the system scales at a cost of between $2,480-4,379 per node per year depending on license agreement length and which OCR module is used.

The cost of the processing platform directly effects the cost of electronic discovery specifically and litigation generally. Let us suppose a given node may process, on average, g gigabytes of native ESI per day. Assuming perfect scalability, the minimum cost of a corpus c gigabytes in size with a deadline of d days can be expressed by first solving the equation $\frac{c}{gd} = n$ to determine how many nodes $\lceil n \rceil$ are required to process c gigabytes of ESI in d days if each node can process g gigabytes per day. The minimum cost, assuming perfect distributed processing and cheapest licensing per annum is $\$6,010 + (\lceil n \rceil - 1) * \$2,480$. Under this formulation, the minimum hardware/software necessary to support processing will vary based on the job size and deadline length which must be supported. Using this we can establish a pricing floor per gigabyte in an ideal, but not obtainable, configuration. By setting $d = 1$ we can then set how many gigabytes we wish to process per day c, and determine the minimum pricing based on how many gigabytes per day each individual node achieves g. Thus, for daily capacity c at per node rate g our minimum cost per gigabyte assuming perfect distribution and processing at 100% utilization over a year would be $\frac{\$6,010+(\lceil \frac{c}{g} \rceil - 1)*\$2,480}{365}$ per gigabyte to break even on the LAW licensing costs.

Software alone will not a process make. The cost of hardware must also be accounted for as it scales equally the node increase assuming, for simplicity, that virtualization is not used and a $1:1$ relationship exists between nodes and servers. A modest rack-mount server with various warranties to account for added

$$\frac{\$6{,}010+(\lceil \frac{c}{g} \rceil -1)*\$2{,}480+(\lceil \frac{c}{g} \rceil *\$1{,}600)}{365}$$

Figure 9.2.1: Ideal Per Unit Daily Capacity Cost Scalability

breakdown costs over five years will run approximately $8,000 or $1,600 per year. Adding this to our equation we arrive at the following formula for node scalability in figure 9.2.1.

9.2.1 Pagination Architecture

What exactly does LAW do when it processes native ESI into a paginated format? LAW utilizes the Tagged Image File Format ("TIFF") for pagination. It does this by creating a virtual printer within Windows' print system[7]. Using the three minimum modules I discussed *supra*, the user creates a new case *via* the Admin module, then imports the native ESI *via* the ED Loader module. With the native data loaded, LAW begins its TIFF conversion process by pushing the native files to the associated native application. The associated native application then prints to the virtual printer, and the virtual printer "prints" to a TIFF file which is then incorporated into the case. Figure 9.2.3 on the next page illustrates what this workflow looks like.

LAW's support system relies on the third party application associated with the file type to process native ESI to TIFF. In some cases manual intervention is required because the printing process cannot be completely automated. The requirement of these proprietary applications also introduces another cost factor. For each LAW processing node, a license of the appropriate application suites must be purchased. We can think of this additional cost as $\sum_{i=1}^{a} cost(i)$ where $cost(i)$ is the license cost for application i. See figure 9.2.2.

$$\frac{1}{365}[\$6,010 + (\lceil \frac{c}{g} \rceil - 1) * \$2,480 + (\lceil \frac{c}{g} \rceil * \$1,600) + \sum_{i=1}^{a} cost(i)]$$

Figure 9.2.2: Ideal Scalability Including Third Party Application Costs

The use of TIFF as the format for pagination is problematic. Firstly, TIFF is an image container whereas the majority of content contained in ESI is textual. Second, TIFF files are large and cause an extreme increase in data set size for the final processed paginated form versus the original native ESI. Third, the conversion to TIFF requires additional steps to capture non-visible information. Finally, conversion to TIFF presents the problem of losing textual information especially text location information within the paginated image. This introduces the need to use OCR to recapture the textual information and its location within the paginated *corpus*.

[7]LAW PreDiscovery is Windows only.

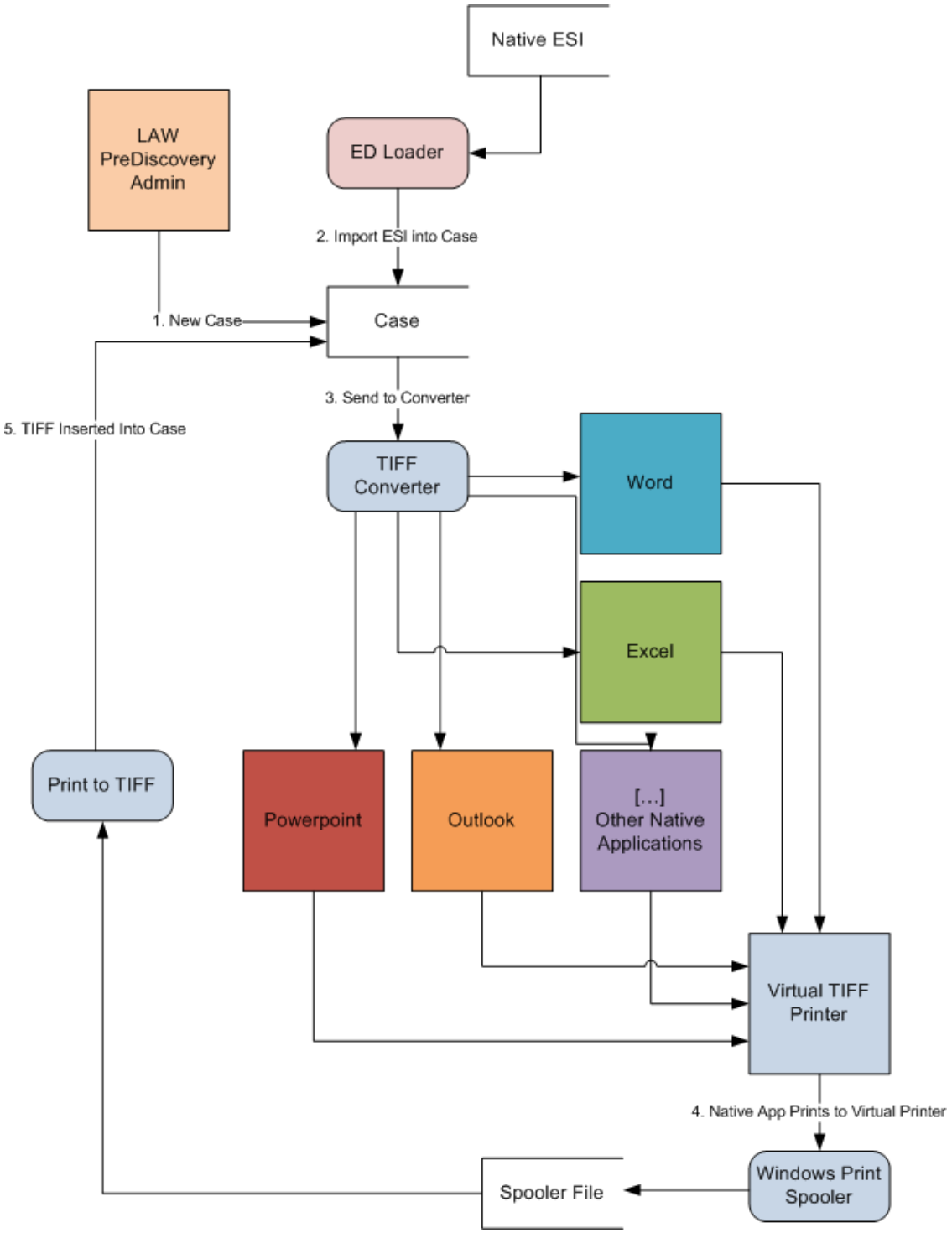


Figure 9.2.3: LAW PreDiscovery TIFF Work Flow

9.3 Pagination with Discere

Two of the main problems with existing approaches are the reliance on virtual printers as a pagination mechanism and using TIFF as the primary paginated format. The virtual printer mechanism introduces an unnecessary bottleneck through the Windows print system. It also introduces certain anomalies with output format particularly e-mails in the commonly used PST file format where the paginated emails contain a 'dummy' user account name as a header due to the way Outlook natively handles printing. TIFF files present a one way problem in that a more text appropriate format such as PDF can be converted to TIFF without issue, but converting TIFF to PDF leaves you with the same file size and OCR problems the TIFF had to begin with. In comparison, PDF files containing primarily textual information are relatively small and closer in size to the native ESI.

In addressing the pagination problem with Discere I chose to use PDF as the primary pagination format for the reasons just explained. PDF has numerous benefits including wide spread support in many applications, small file sizes for primarily textual documents, and easy conversion to various image formats if required. In lieu of using a virtual printer architecture, I utilize the native associated application itself where possible to do the conversion and alternatively build the paginated PDF output programmatically where the native application will not support direct conversion.

Where existing approaches utilize proprietary applications, I instead use open source components. The substitution of open source components produces numerous benefits both cost focused and functionality focused. Cost wise, open source components remove the need for additional software licenses per node. Technologically, open source components present the ability to interface directly with the application instead of relying on a specific vendor to allow for programatic incorporation. Open source also allows Discere to leverage significant development investment from the open source ecosystem into creating a robust processing application without reimplementing rendering components for each file type; additionally, Discere will benefit from the continued development of the open source components it relies on when those components are updated.

9.3.1 Email

In litigation, e-mail is predominantly found in Personal Storage Table ("PST") format files generated either from Outlook as a local mail-store or as an export from an Exchange server instance. PST files prior to the 2003 switch to 64 bit addressing were limited to 2GB, while post 2003 files have a much larger maximum (20-50GB by default.) By using an appropriate library implementation[22] the internal content and folder structure of the PST file can be navigated without the presence of an Outlook installation. In existing

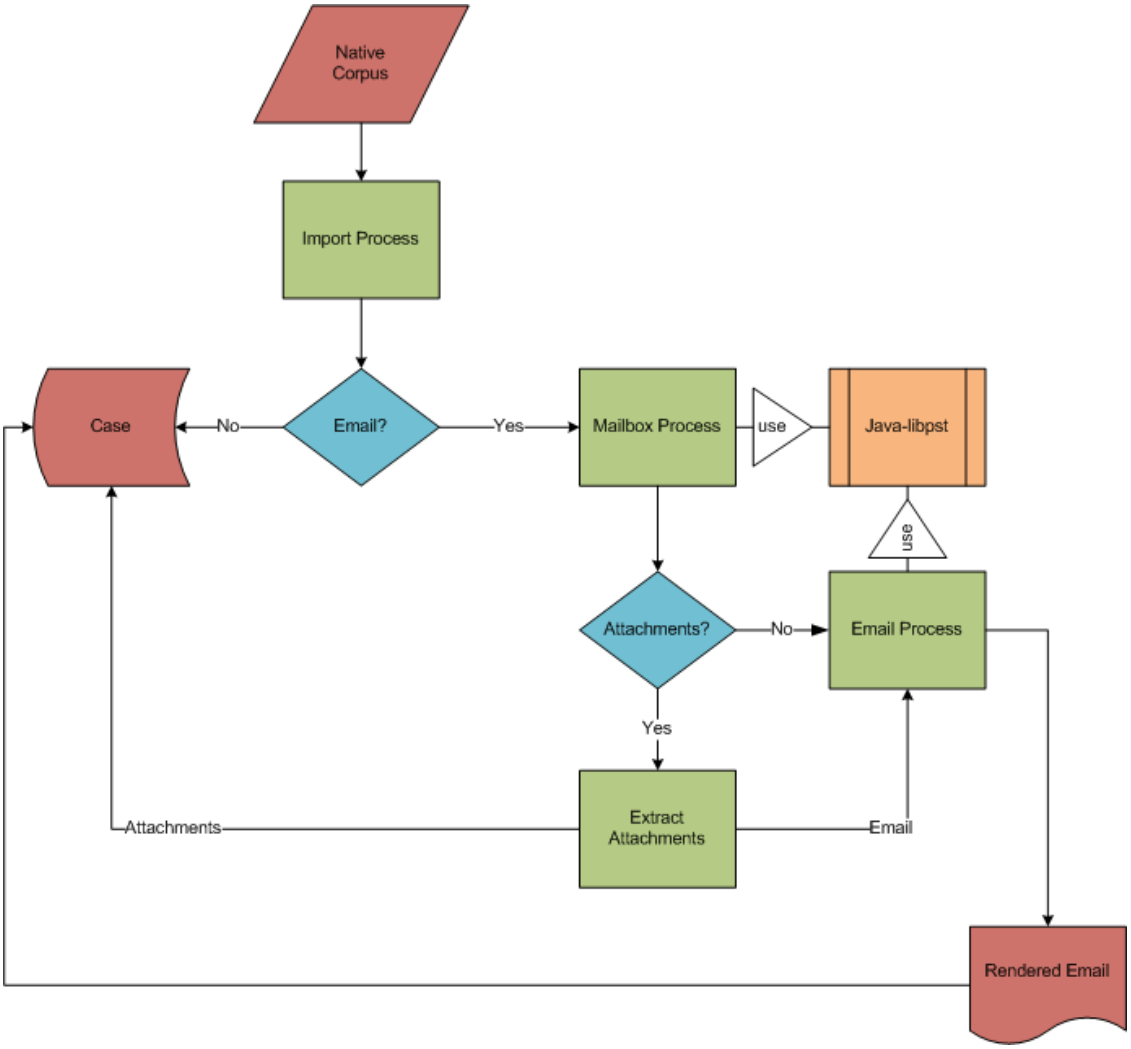


Figure 9.3.1: File / Email Import Work Flow

industry approaches the utilization of Outlook for PST processing restricts such tools to the Windows platform. Reliance on Outlook also introduces formatting artifacts in that Outlook produces the user profile name on the top of each email when printed, and the use of dummy profiles for discovery processing causes these artifacts to appear regardless of email source.

PST files contain emails, appointments, contacts, and other information. The contents, such as emails, may themselves be thought of as containers as is the case when emails have attachments. Processing PST files requires iterating through the contents while preserving parent-child relationships, then treating each contained item as a potential container (see figure 9.3.1 on the previous page). Emails consist of more than just visual elements; in rendering a paginated email the visual cues of contained data sources (such as attachments) must be preserved, the individual child attachments must be extracted for further processing, and the data sources must be parsed into the human readable fields (to, from, cc, bcc, subject, date, body, etc).

MSG files must also be handled in a similar way to PST files in that they to are email containers, though of a single message, but may contain other MSG files as well as other attachments. In a similar vein to PST files, existing libraries provide access to the underlying data structures in the MSG file so they can be rendered more efficiently than relying on Outlook.[3]

9.3.2 Office Documents, Images, & Drawings

Office Document, file types generally associated originally with Microsoft Office and now more widely supported by various productivity suites, are the bulk of file types typically handled in discovery beside emails. In many ways, email and office documents go hand in hand as such files are plentiful in email *corpora* as attachments. There are substantial differences in their formats from the pre-2007 binary formats in old Object Linking and Embedding structure ("OLE"), the 2007/2010 XML based format, and the "open formats" such as Open Document Format for Office Applications (ODF) first created by the OASIS consortium and later adopted as an ISO standard.

Originally, Discere relied on OpenOffice for Office Documents via the Universal Network Objects API. Subsequently, Discere incorporated an open source library, JODConverter, which is centered around controlling OpenOffice processes for conversion purposes.[23] Later still, with the acquisition of Sun by Oracle, the OpenOffice project forked into OpenOffice (now an Apache project) and LibreOffice (the fork). Discere will work with either project currently, and in some cases utilizes both due to one or the other not being able to handle a particular file due to its content and/or an existing bug. With each release of the projects new features become supported, and performance improvements are realized.

Libre/Open office also provide support for certain file types which are not traditional Office Documents, including certain vector image formats such as DXF[8] files. While individual image files generally are better handled directly, the support for DXF is very important for data sets from engineering companies where they are prevalent.

9.3.3 ASCII or Unicode Text

Text itself is easily handled through the PDF libraries themselves.[2, 19, 21] By handling the text as a direct addition, it remains searchable in contrast to approaches which convert such information into images. This both decreases the resulting file size, and removes errors associated with later OCR used to recapture the lost textual information.

9.3.4 Binary Files

Binary Files which are not presently supported, or where the format does not lend itself to pagination are handled in three ways. First, the Apache TIKA library[5] provides support for extracting textual and metadata information from a wide variety of file types. Many of its supported file types overlap with those better supported by other parts of the system, and in such cases is only used as a fall back for failed conversions.

Where textual data cannot be extracted from the binary file, an extraction of ASCII string data is attempted. As a final failure handler, placeholder files are used for files not amenable to either TIKA or String Extraction.

9.3.5 Archives

Archive files, or file containers contain other files, but are not themselves clearly suitable for meaningful pagination. I have chosen to render these files by constructing PDF listings of their available metadata, and the files contained within them. The parent/child information is recorded so that contained files can be located from a given container entry as well as allowing the container to be located when encountering a specific file of interest. Archive support for a variety of formats is available as part of the base Java API.

[8] DXF files are commonly exports of AutoCAD formats that are more widely supported due to the proprietary nature of the .DWG format.

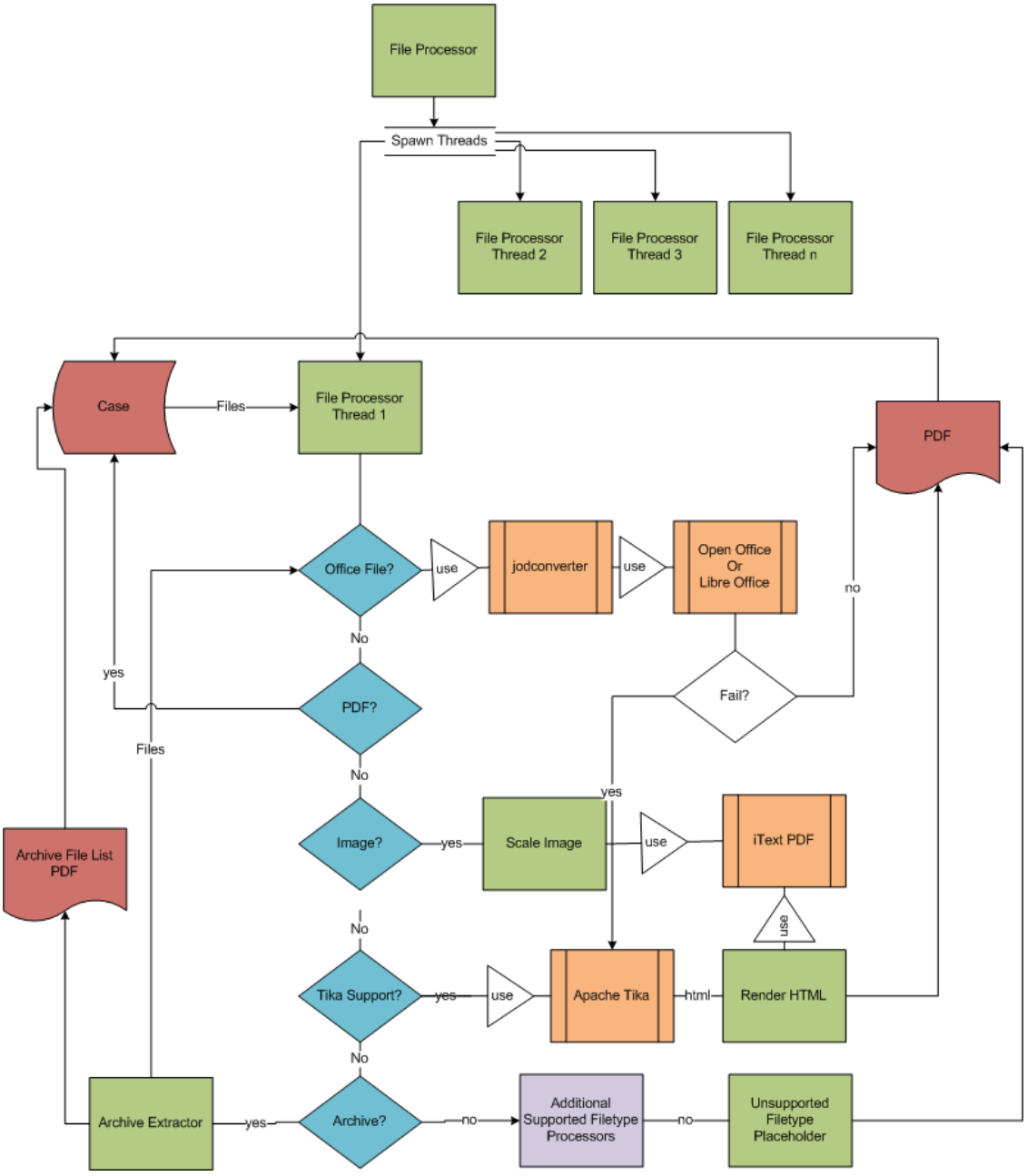


Figure 9.3.2: File Conversion Work Flow

9.4 New Project Interface

9.4.1 Creation

Project creation is straight forward from the user perspective. After selecting the new project menu item, the user is prompted for a location to create the new project's root directory (see figure 9.4.1 on the following page). The root directory is created with a number of subdirectories for different aspects and/or stages of the process. The *native* subdirectory holds the imported original or NATIVE files. The *pdf* subdirectory holds the rendered / paginated versions of the NATIVE files in PDF format after processing. The *tiff* subdirectory holds tiff versions of the PDF renderings; the rendering to tiff files is an optional step included for legacy compatibility purposes. The *index* subdirectory holds the Lucene index data extracted from the documents and is used in the search / review components of the tool; a new Lucene index is created within the subdirectory. The *db* subdirectory holds the project database information, and a new instance of an Apache Derby database is created with project tables. .Finally, a *temp* subdirectory is created for temporary file creation used throughout the tool.

9.4.2 Import

Importing files into the system is done by selecting the Import Files option from the Project menu (see figure 9.4.2 on the next page) then selecting the source data from the file selection prompt (see figure 9.4.3 on page 115). In the test example used for the illustrative figures in this chapter, a subset of the Enron Corpus is used consisting of 6 PST files ranging in size from 25 MB to 69 MB for a total size 325 MB (see figure 9.4.4 on page 115).

Once the source data is selected, the system copies the native files to the *native* subdirectory of the project root. A progress bar is displayed for the user to monitor, but the progress indicator aggregates progress from multiple sources: the file copy itself, hash value calculations, creation of entries in the database, and special handling procedures for container files and email files.

Container files are handled through a recursive procedure. When a container file is encountered, its contents are extracted to the *native* directory like any other file and the parent-child relationship is recorded in the project database. Email files which are also container files, such as PST files, are handled in a similar way with some deviation. Email itself contained in such a container does not exist as a discrete file, but rather as an amalgamation of information representing the original email. When an email is encountered in such a state, a paginated rendering of the email is made into a PDF format then added to the *native* directory with the same parent-child relationship noted. As stated, this process is recursive and an email

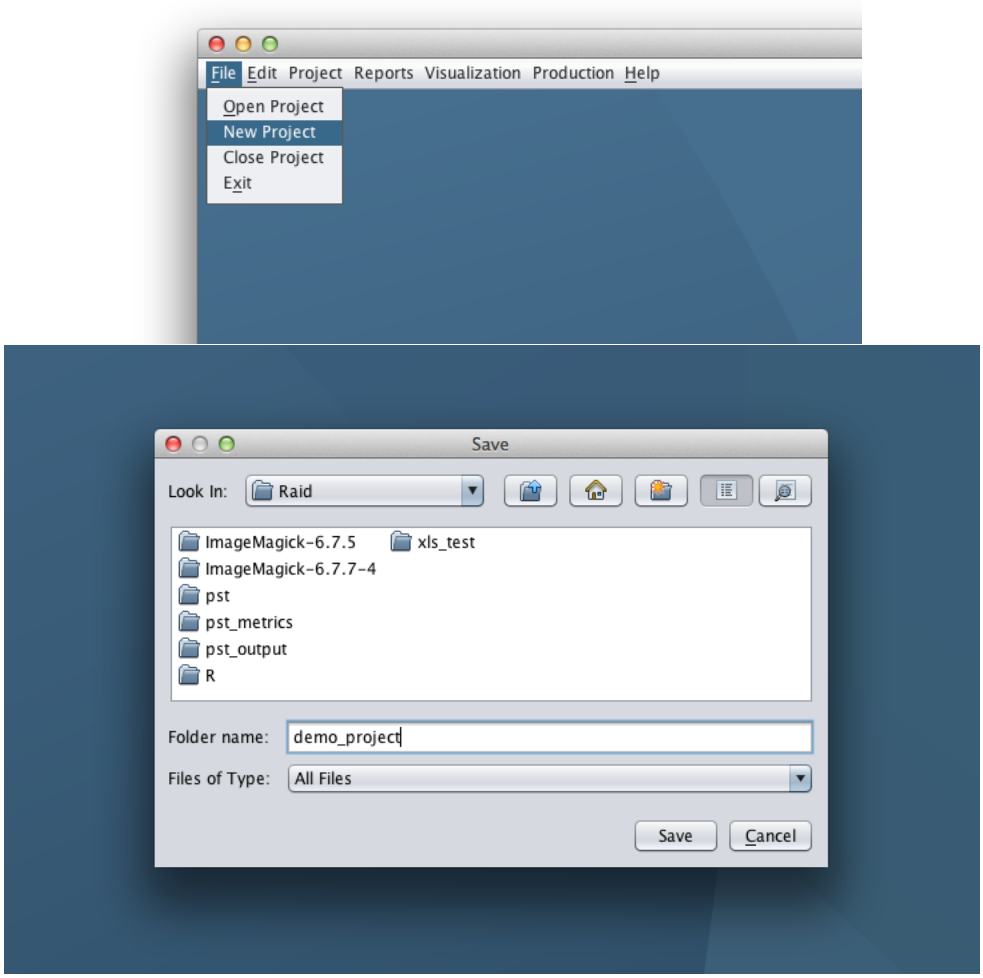


Figure 9.4.1: New Project

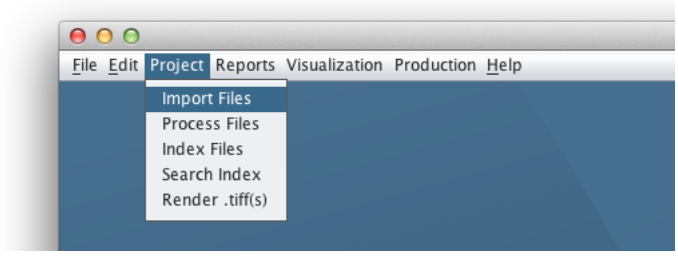


Figure 9.4.2: Import Data

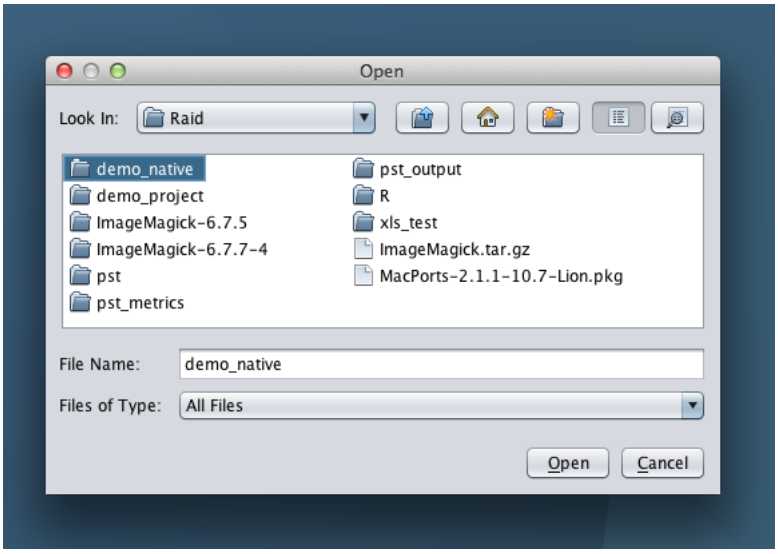


Figure 9.4.3: Data Selection

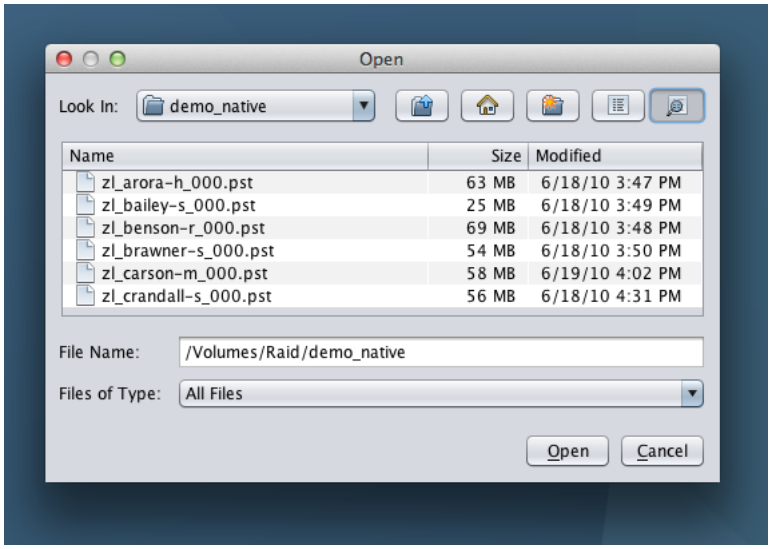


Figure 9.4.4: Data Selection Details

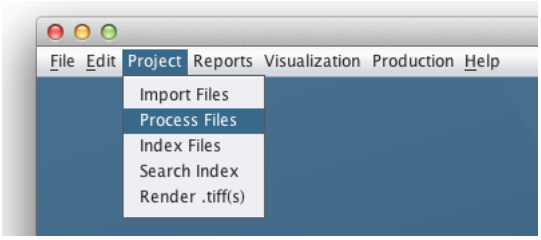


Figure 9.4.5: Process

might also be a container with its attachments processed in the same manner. The end result, in such a recursive chain, is that the relationship of a given attachment can be traced to the containing email and so forth until the root container, the PST file, is reached.

When all the native files are copied, the records created, and any required renderings completed, the progress dialog displays a 100% complete message indicating additional actions are permitted.

9.4.3 Processing

The *processing*[9] step involves the conversion of all original files into paginated format. The process employed here differs substantially to the standard procedures employed elsewhere in that the conversion goes to PDF rather than TIFF. The process is started by selecting the Process Files option in the Project menu (see figure 9.4.5) which starts the file processor thread.

The file processor is an instance of ThreadedFileProcessor which extends SwingWorker. When instantiated, the file processor queries the project database to determine the number of unconverted native files remaining. It then determines, from the system, the number of available CPUs and spawns an equal number of worker threads allocating native files for conversion to each thread in chunks, and distributing further chunks as the worker threads finish. The progress of the conversion task is displayed in a progress bar dialog.

The specifics of conversion for a specific file depends on its type (see section 9.3 on page 108). Generally, the majority of files handled in discovery involve textual documents, spreadsheets, and other "Office" formats. Handling of these types occurs through either OpenOffice or LibreOffice. Because the process is threaded to fully utilize all available CPU resources, the system will necessarily spawn a number of instances of the given document suite equal to the number of worker threads being used which are, in turn, equal to the number of available CPUs (see figure 9.4.6 on the following page). When a file fails to render because of an error or where the file type is not supported or not subject to pagination, the system notes the failure in the project database and generates a placeholder PDF file.

When conversion is complete, all instances of conversion tools (including instances of OpenOffice or

[9]The processing step refers to the conversion aspect of processing stages referenced elsewhere.[41]

Figure 9.4.6: Process Instances

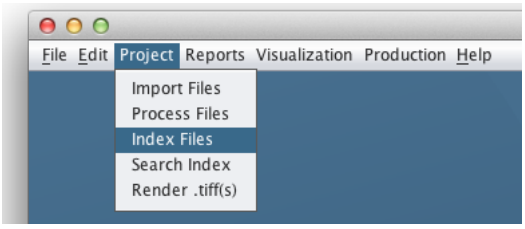


Figure 9.4.7: Index

LibreOffice) are destroyed and the progress dialog indicates 100% complete indicating additional actions are permitted.

9.4.4 Index

Indexing is handled in a straight forward manner. The indexer is an instance of ThreadedFileIndexer which, like ThreadedFileProcessor, is an instance of SwingWorker. The indexer utilizes the Lucene project[42]. Each thread will iterate over a chunk of the files extracting text directly from the PDF or, if not paginated, from the native file using either a supported parser (see section 9.3 on page 108) or a simple string extraction.

Pursuant to the Lucene project best practices, the indexer instantiates a single IndexWriter and provides it to each spawned thread. The progress is updated as the threads complete their assigned chunks. When the index threads are complete the system optimizes the index before committing any remaining changes. At this point, the index is complete and ready to be searched.

9.5 Error handling

The presence of errors in the conversion process is both normal and expected. Files may fail to convert for a number of reasons, but one of the most common is the conversion process for a specific file exceeded the timeout allowed for. Because bulk conversions is both CPU and I/O intensive, the conversion process itself fairly cumbersome, and the nature of the system here is multi-threaded, a particular file may simply take too long while other resources are constrained or may be an outlier file which simply requires more than the allowed 30 second time out to complete. This section shows the handling procedures for errored file

FileId	Error Type	Message	Notes	Entered
4453	1	task failed	File Name: 4453.CURVE_REVIEW_BECO_AUGUST_2001.ppt	2012-06-08 13:06:51.457
354	1	task did not complete within timeout	File Name: 354.Vol Book Daily Market.xls	2012-06-08 13:07:57.99
541	1	task failed	File Name: 541.PJM Price Curves.xls	2012-06-08 13:09:37.76
11636	1	task did not complete within timeout	File Name: 11636.12-10-2001.xls	2012-06-08 13:14:14.127
9735	1	task did not complete within timeout	File Name: 9735.EOL_All_Swaps.xls	2012-06-08 13:14:54.362
12720	1	conversion failed	File Name: 12720.GTAFacilitiesinRTOWest.xls	2012-06-08 13:20:34.909
12733	1	conversion failed	File Name: 12733.Pricing-alt-final.xls	2012-06-08 13:20:39.374
12970	1	conversion failed	File Name: 12970.Pricing-alt-final.xls	2012-06-08 13:21:06.405

Figure 9.5.1: Error Report

conversions.

9.5.1 Error Report

The error report shows which files failed in the conversion process, when it happened, and what file id represents the file (see figure 9.5.1). Files which fail conversion will still have placeholder PDF files and will be represented in the search index. The first entry in figure 9.5.1 for FileId 4453 is for a Power Point presentation (.ppt) which we can query for in the index by searching $FileID : 4453$ which isolates the search to the FileID field using the Lucene query syntax (see figure 9.5.2 on the next page)[42]. Double clicking on the returned result will bring up the review tool displaying the placeholder file associated with the index entry (see figure 9.5.3 on the following page)

9.5.2 Reprocess

Often, the error in conversion is a result of a time out and will usually not reoccur when either processed individually, *id est* when the system is not under heavy load, or when the time out is increased. The review tool window has a REPROCESS button in the lower left corner (see figure 9.5.4 on page 120). When selected, the reprocess option will increase the timeout to 60 seconds and attempt to run the conversion for the given file again; if successful it replaces the placeholder file, reindexes the new pdf, and displays the updated file in the reviewer tool (see figure 9.5.4 on page 120). If reprocessing fails, the next stage in error handling is to manually replace the rendered file.

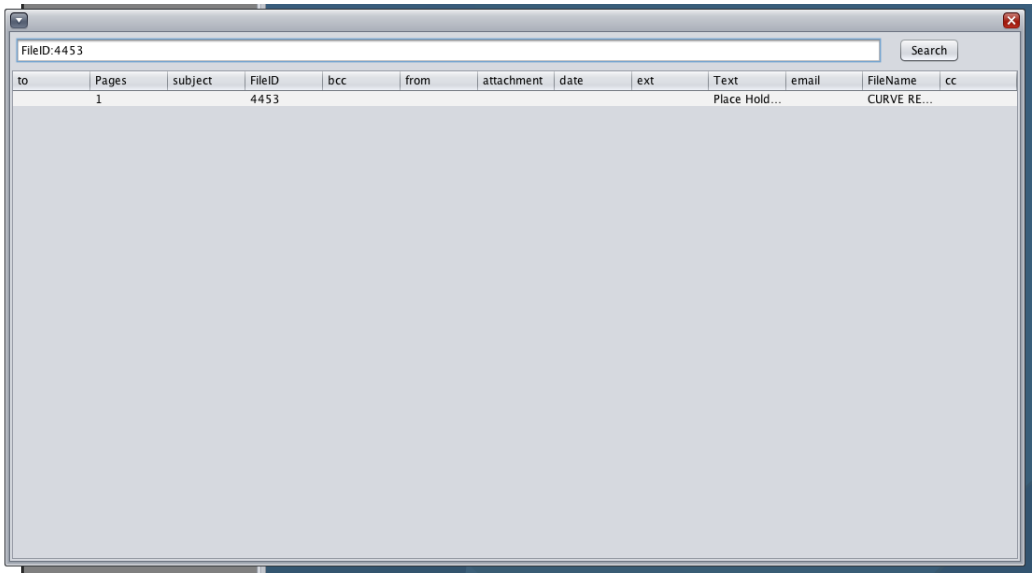


Figure 9.5.2: Error Lookup

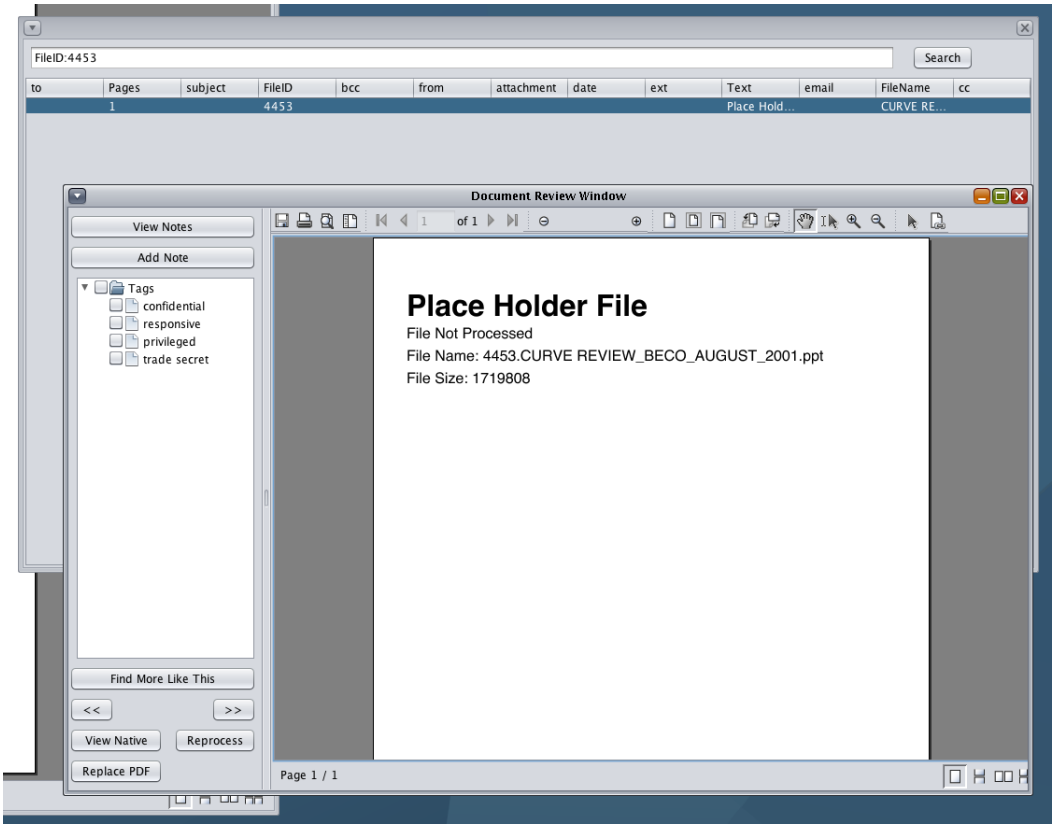


Figure 9.5.3: Placeholder File

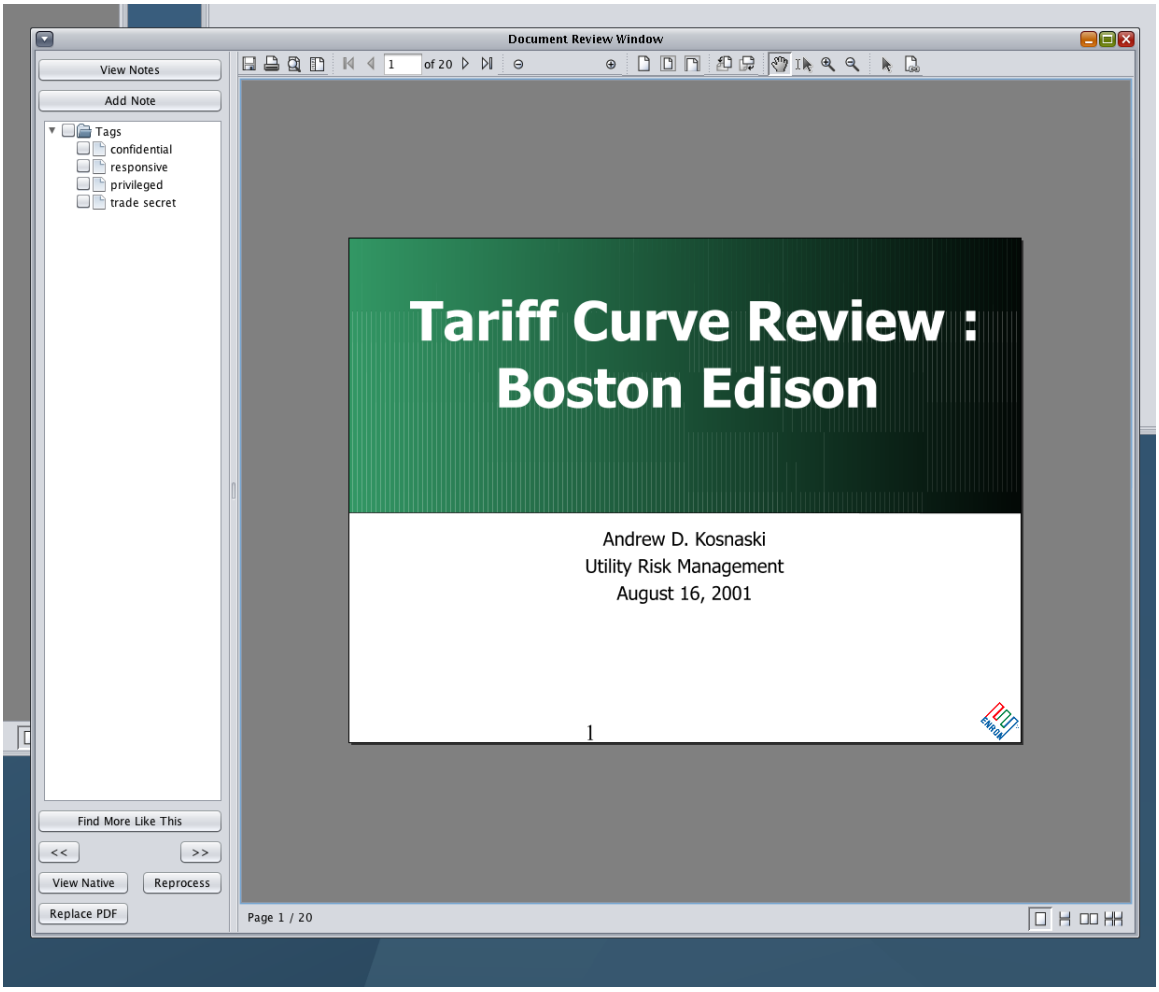


Figure 9.5.4: Reprocess File

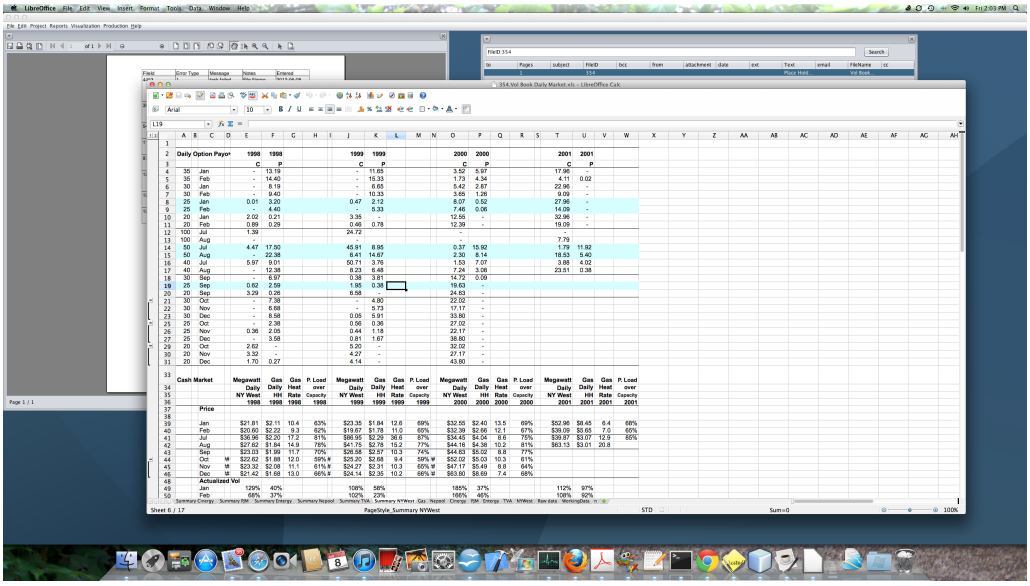


Figure 9.5.5: Open Native

9.5.3 View Native / Manual Replace

Manual replacement of the rendered file is useful for when reprocessing fails to properly convert the file, for adding PDF conversions for file types not supported by the system, and for updating a rendered conversion with a better formatted version. The VIEW NATIVE button in the review tool window will launch the application associated with the given file's type; in the case of File 354, this launches Libre Office to open the spreadsheet (see figure 9.5.5). The spreadsheet in question is abnormally complex (see the number of worksheets present) and each sheet is fairly long and wide which explains the timeout failure in the first pass. While reprocess would work on this file, this is a file for which it may be desirable to manually convert so as to preserve a human readable format.

Typically, with such spreadsheets not formatted for printing, the spreadsheet must be converted carefully in order to keep meaningful information on the same page when converted. Since this document has so many worksheets with varying information / dimensions an automated conversion is highly unlikely, even using landscape with 11x17 page size, to preserve the information in the same way the sheets display it. If, for example, columns A-L on the active sheet are cut from the remaining columns, the information in column O becomes difficult to pair with the information in column A because, in a PDF, the two sheets are not consecutive.

Using the built-in PDF export function, the spreadsheet can be saved as a PDF (see figure 9.5.6 on the following page) to the local disk. Once exported, selecting the REPLACE PDF button from the reviewer tool window will prompt for the PDF location. Once selected, the system replaces the PDF with the specified

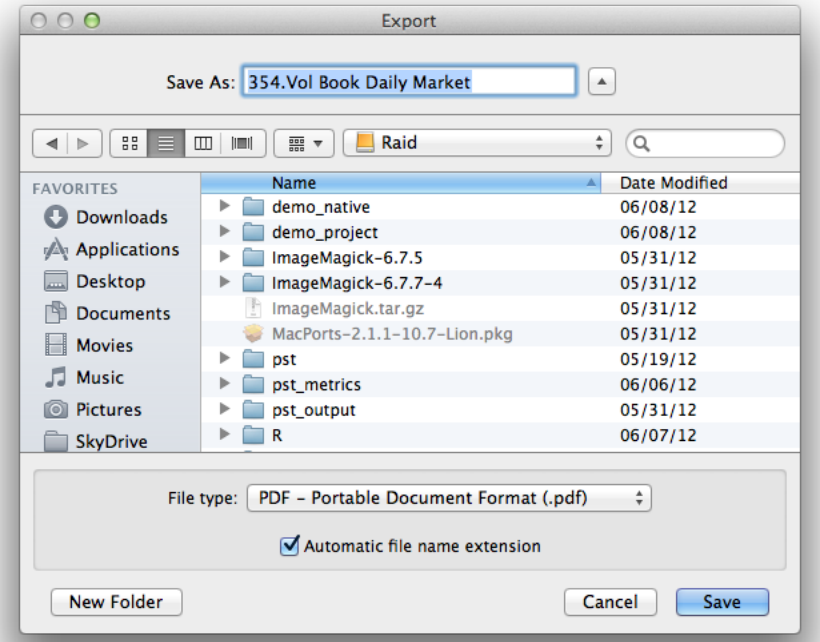


Figure 9.5.6: Manual PDF Export

one, updates the index, and displays the new PDF in the reviewer tool window (see figure 9.5.7 on the next page).

9.6 Searching

Searching in Discere is handled using the Lucene engine which provides a robust, but powerful search query language. Searching with standard boolean logic is supported over the default text field, but the system also allows for searching specific fields. File types such as emails contain information (to, from, subject) which do not appear in normal documents, but which may be useful in forming searches or filtering results. A search for fraud, for example, brings up a number of emails in a particular user's mail-store (see figure 9.6.1 on the following page).

From such results, one could add additional search terms/fields to filter the results or explore the individual items to see where the terms are appearing in the document. By double clicking on one of the results, the review tool window appears with the document in question. By expanding the search option in the PDF display, the text can be searched for to ascertain whether this is a relevant document (see figure 9.6.2 on page 124). This is an example of the ordinary search system which most people would find familiar. Discere

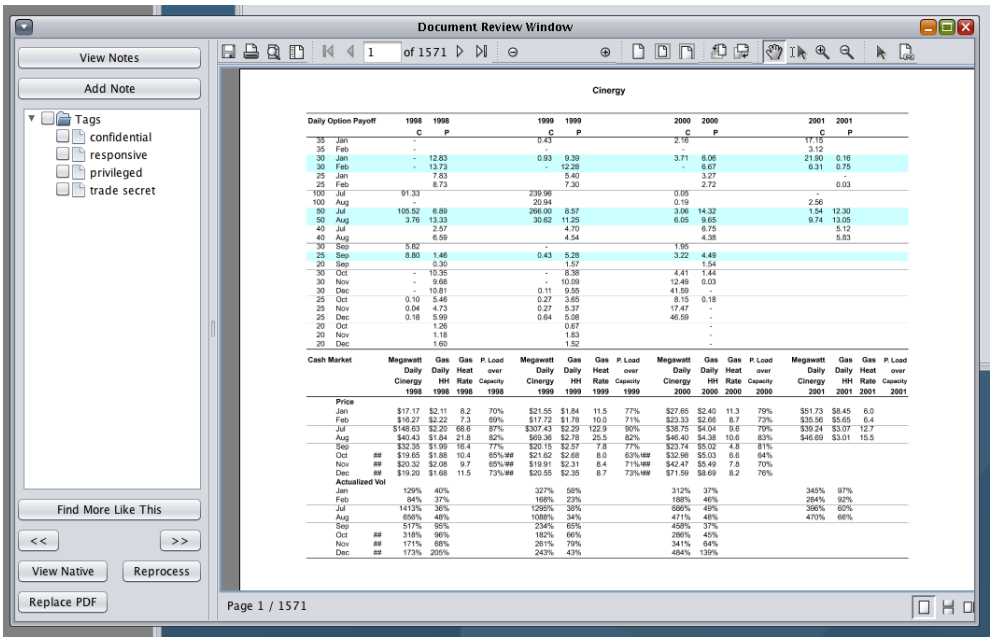


Figure 9.5.7: Replace PDF

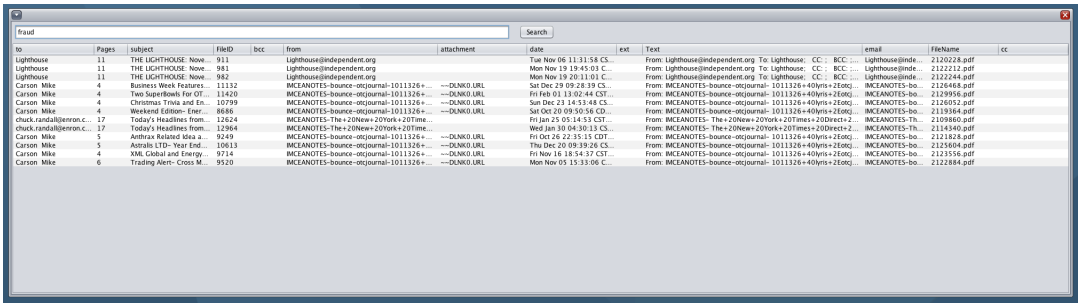


Figure 9.6.1: Search - Fraud

also supports clustering of search results using the *Carrot*^{2}[7] engine (see examples Dealbench on this page, Vasquez on page 125).

Dealbench In the Arora email archive from the Enron dataset[12], mention is made of a system called 'DealBench' used by Enron. A simple search for dealbench produces results from a number of emails and attachments; opening one of the emails shows a conversation thread mentioning the use of DealBench for a new project (see figure 9.6.3 on page 125). Exploring the search results linearly is time consuming especially when the searcher is attempting to locate specific information related to the search, but not all information produced by the search. Using a clustered search on the same dealbench query (see figure 9.6.4 on page 126) partitions the search results into generated clusters based on keywords displayed in the left panel.

While exploring the linear search set to learn more about dealbench, we might find information showing that dealbench is a business (see figure 9.6.5 on page 126). Browsing through the clusters, the WORKING,

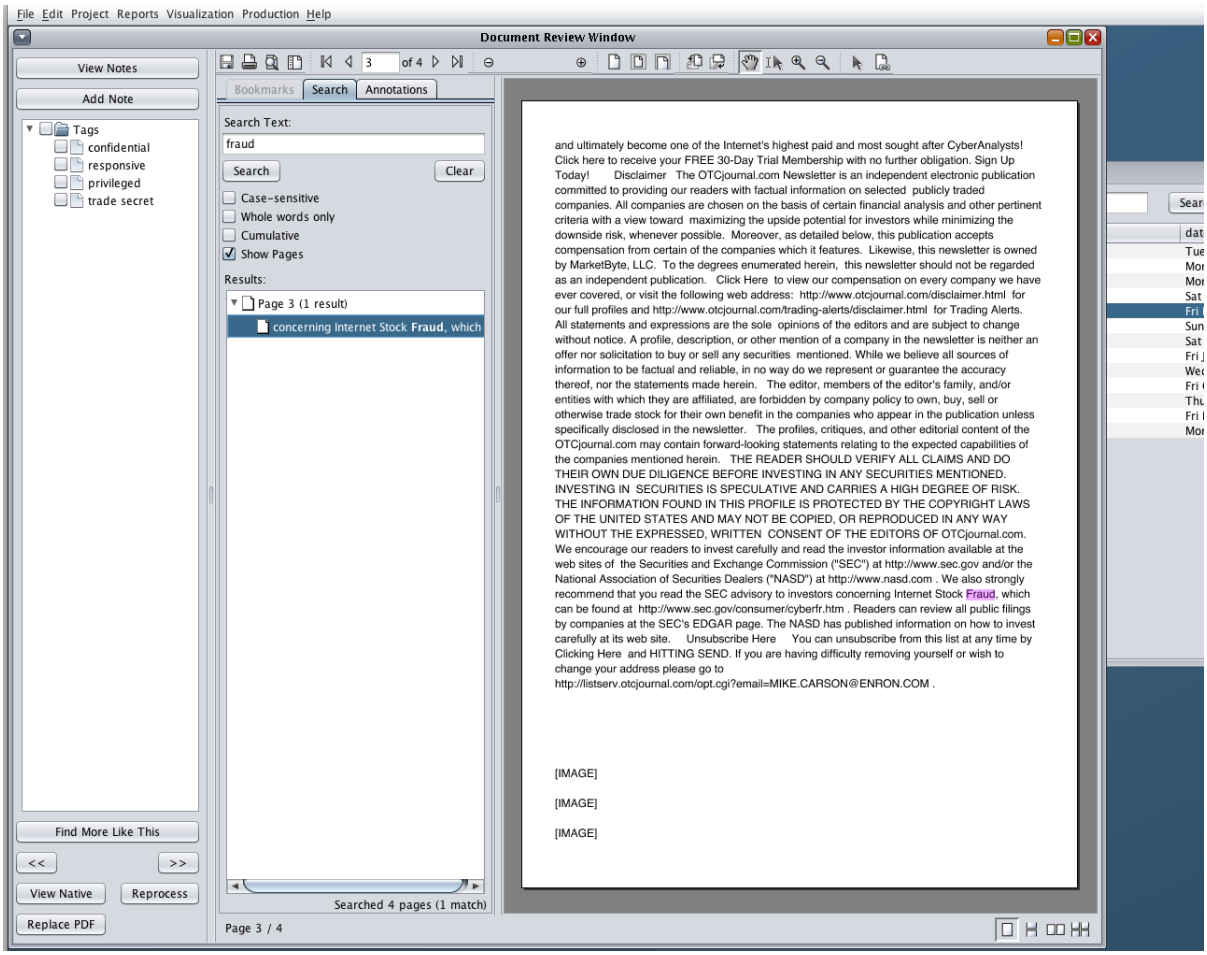


Figure 9.6.2: Search - Fraud / PDF Search Function

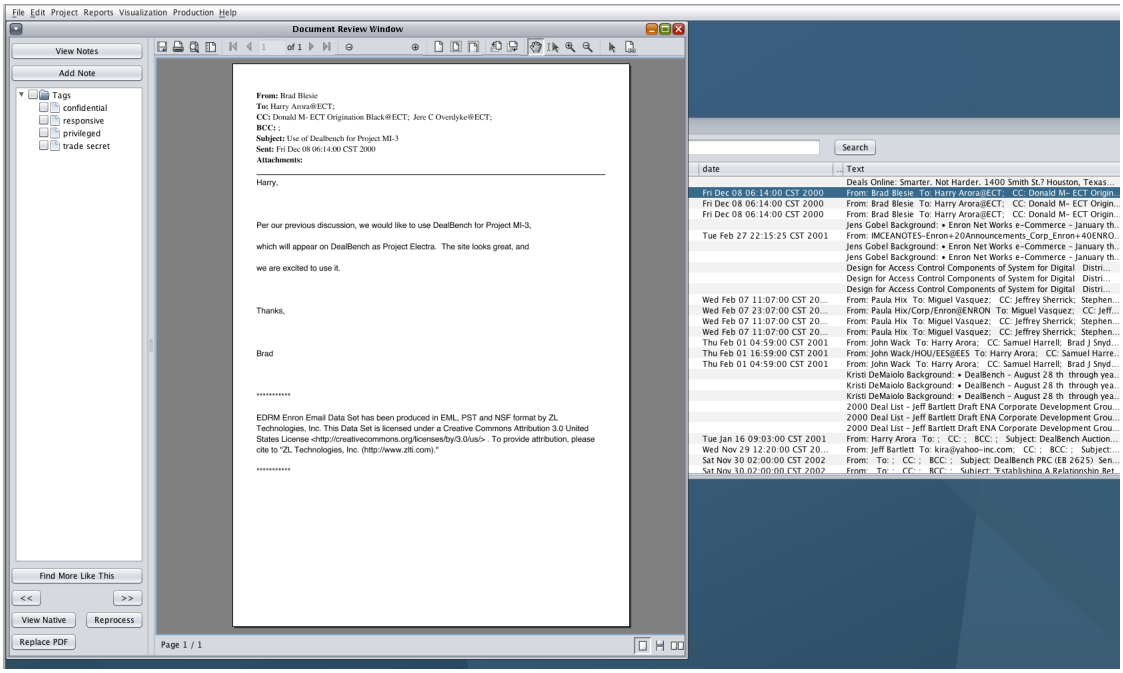


Figure 9.6.3: Search - Dealbench

Summary, Work cluster we can see several instances of a business plan which would likely give us more information on what DealBench, LLC is and/or does; this in turn will help determine if it is relevant to the litigation at hand (see figure 9.6.6).

If Dealbench is relevant, we might also look at the other clusters generated by the search. One of the clusters contains names. Looking in the Kristi Demaiolo, Miguel Vasquez cluster we find employee reviews and other information related to Dealbench with Miguel Vasquez as one of the reviewers. The clustering approach quickly allows us to ascertain certain information about the topic we are searching for, and to generate new search queries based on clusters we find relevant. In this case, we may be inclined to search for information related to Vasquez (see example Vasquez on the current page).

Vasquez Changing our search term to vasquez we obtain a new set of results and clusters to work with (see figure 9.6.8 on page 128). With some searching around we find that despite Arora's leaving Dealbench (see figure 9.6.5 on the next page) Arora still kept contact with individuals such as Vasquez (see figure 9.6.9 on page 128). The ease with which cluster searching allows a user to drill down from single word searches to relevant, related information and exchanges makes it a superior search method over purely linear review of searches. This becomes more evident as the data set size increases making linear review less feasible especially where time constraints exist.

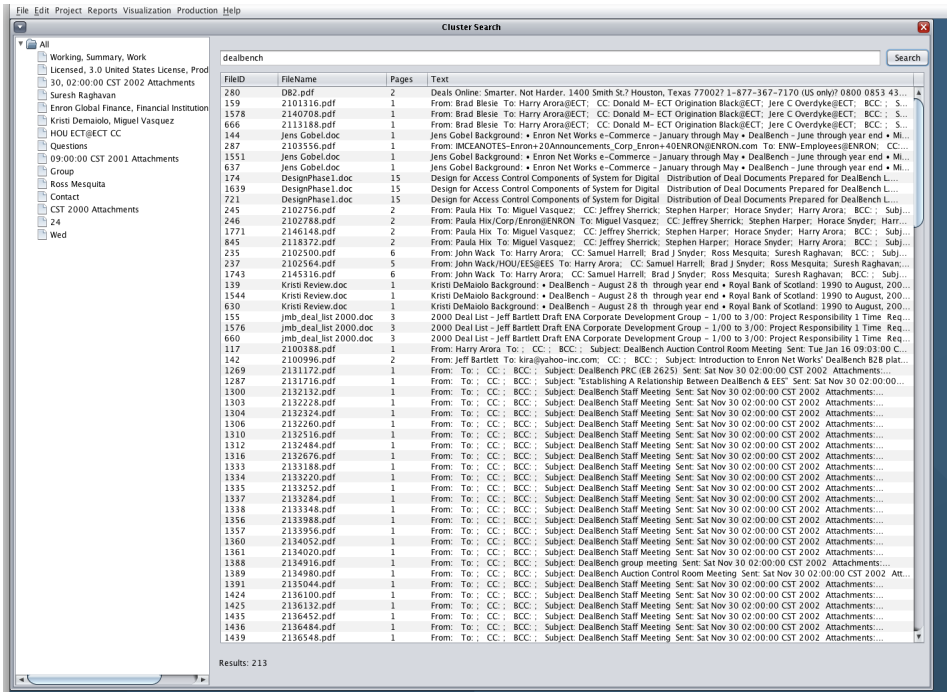


Figure 9.6.4: Search - Dealbench Cluster

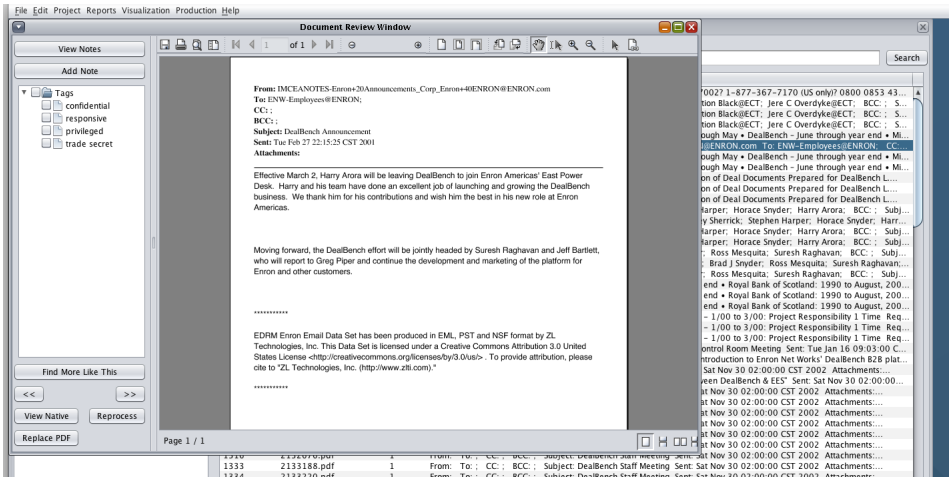


Figure 9.6.5: Search - Dealbench Cluster, Arora Leaving

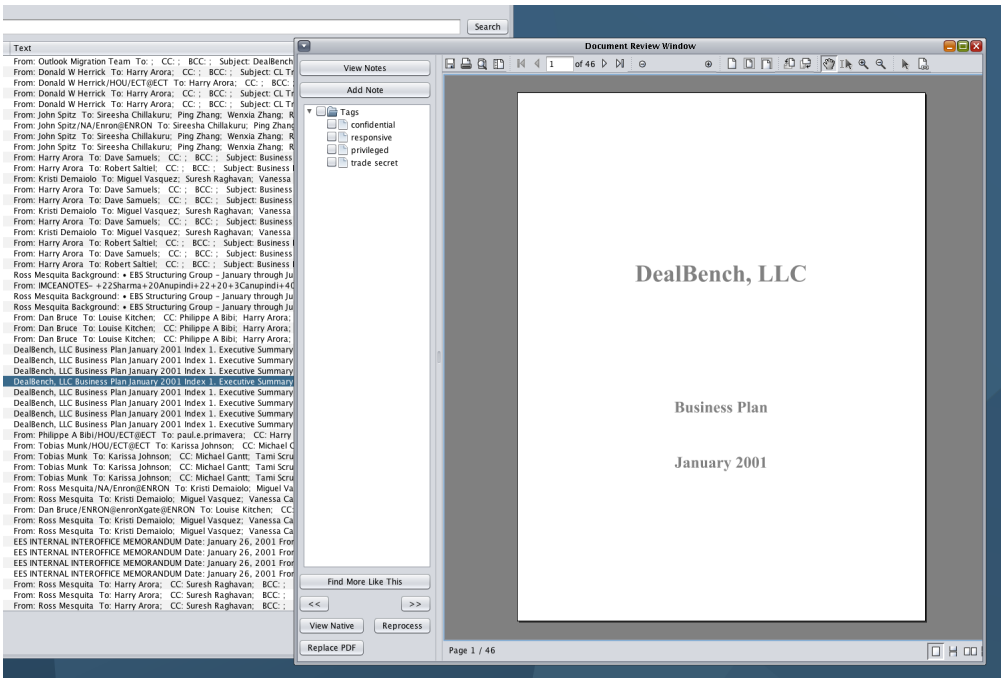


Figure 9.6.6: Search - Dealbench Cluster, Business Plan

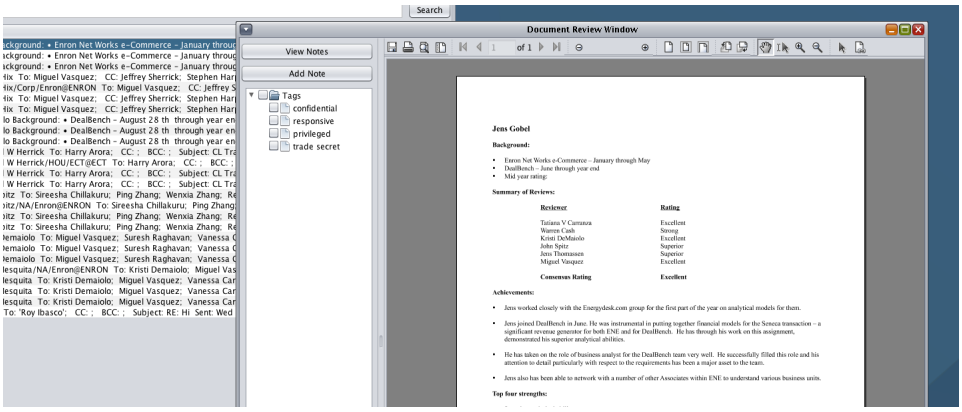


Figure 9.6.7: Search - Dealbench Cluster, Employee Review

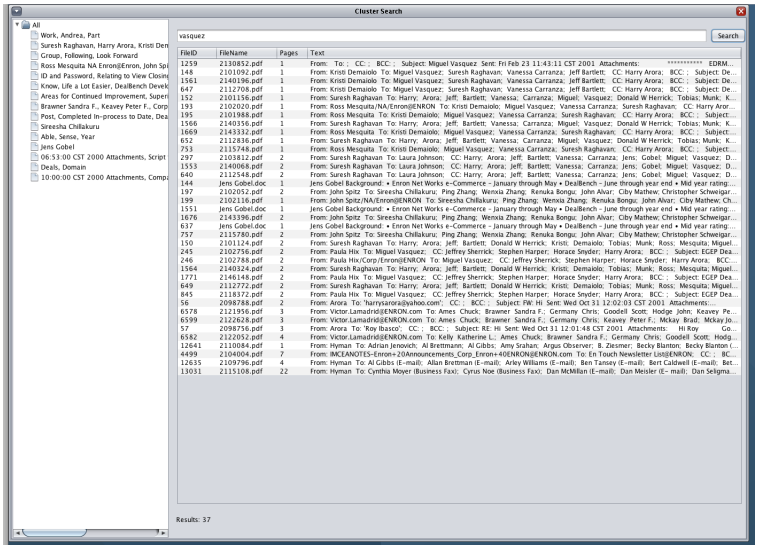


Figure 9.6.8: Search - Vasquez Cluster

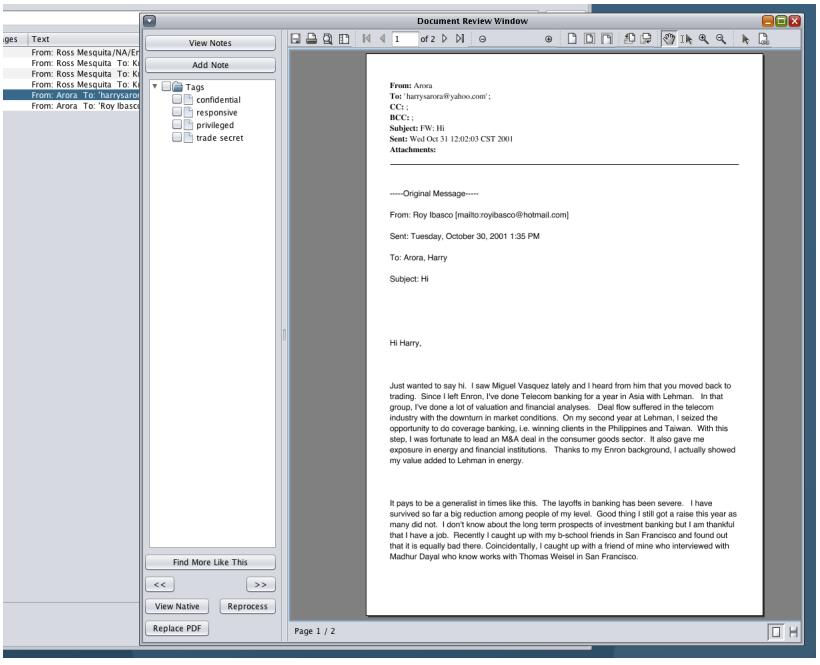


Figure 9.6.9: Search - Vasquez Cluster, Arora

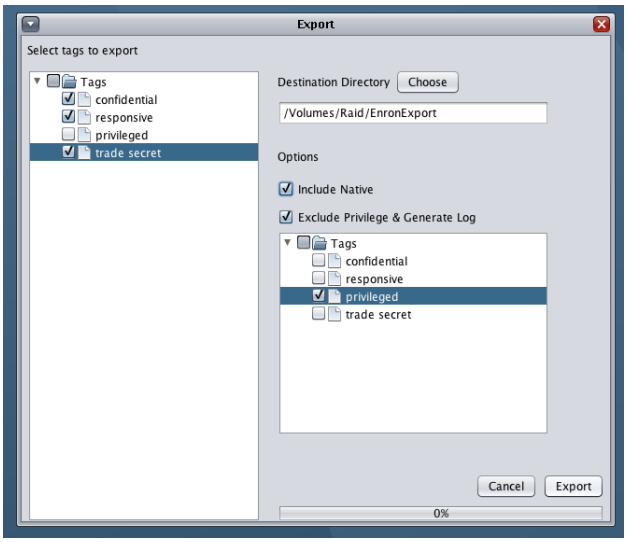


Figure 9.7.1: Export Settings

9.7 Tags and Export

In the figures of section 9.6 on page 122, the review tool window contained a list of tags with checkboxes to apply or remove them from a given document. The default set used in Discere consists of CONFIDENTIAL, RESPONSIVE, PRIVILEGED, and TRADE SECRET; tags can be added or removed, but the default set is illustrative for this section. As the corpus is reviewed, the reviewer can set tags which describe the document at hand. In Discere, tagging a document also tags identical versions of the document to avoid duplicative efforts. If the litigation involved Dealbench, LLC, hypothetically, then the document in figure 9.6.6 on page 127 (the Dealbench, LLC business plan) would be relevant and tagged as responsive, but the employee review for Jens Gobel (see figure 9.6.7 on page 127) might not be, and further might be tagged as confidential to ensure exclusion. The tagging process allows the user to describe the document in terms of the tags, with an eye to later exporting certain tags while excluding others. This, in essence, is the document review process - categorizing documents into a taxonomy in relation to litigation.

When review is complete, the document set can be exported to provide to another party. The export window allows the user to select which tags to export, where to export the documents, whether to include the native files, and whether specific tags should be excluded. In the example in figure 9.7.1, all CONFIDENTIAL, RESPONSIVE and TRADE SECRET documents will be produced, but any of those documents tagged as PRIVILEGED will be excluded; the resulting export will be saved to /Volumes/Raid/EnronExport.

When the export is complete (see figure 9.7.2 on the next page) the output destination contains a native directory, a pdf directory, and a priv_log.txt file (see figure 9.7.3 on page 131). The priv_log.txt file contains a list of files excluded from the export along with the tag which triggered the exclusion (see figure 9.7.4 on the

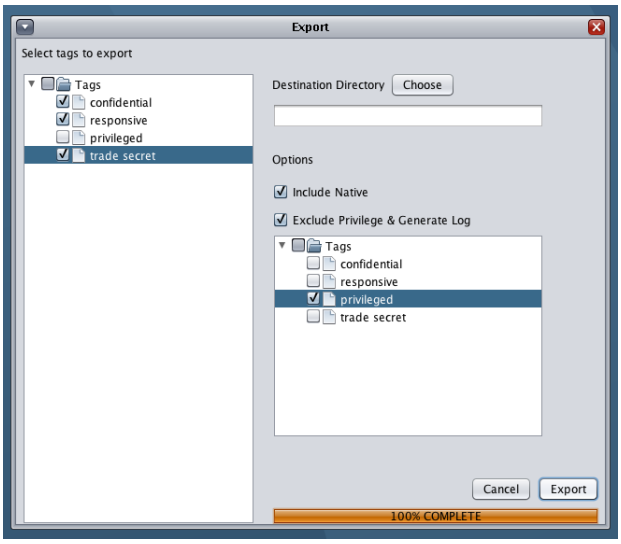


Figure 9.7.2: Export Complete

next page). The output can be provided to another party in response, for example, to a discovery request.

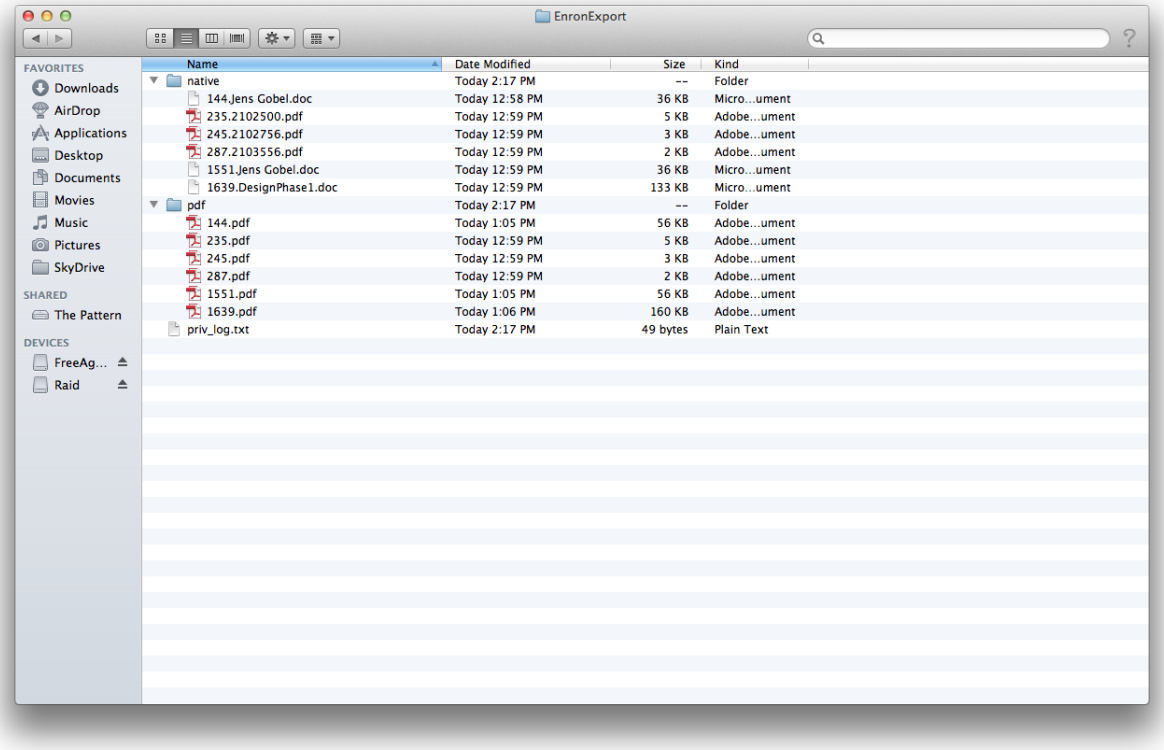


Figure 9.7.3: Export Directory

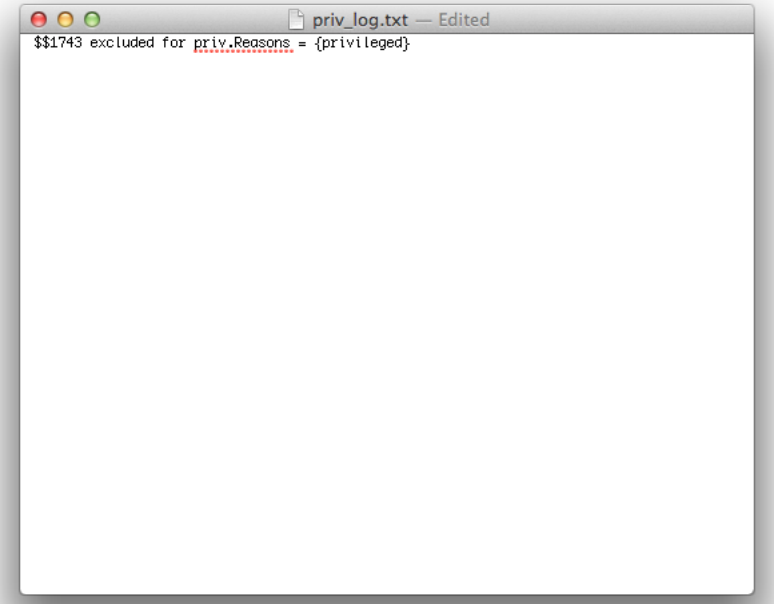


Figure 9.7.4: Export Privileged Log

Chapter 10

Performance

This chapter presents and describes performance metrics for the conversion of various common file formats. Some file format types, such as .doc or .xls, are prevalent in far greater numbers than other formats such as .html. In each case, the data points are plotted in 2 dimensional euclidean space. Because of the data point density and performance variance in the file formats, *id est* some formats are non-linear in a time/size comparison, the data is subjected to Linear Least Squares Regression in R with the resulting line plotted in red.[49]

10.1 Aggregate Processing Rates

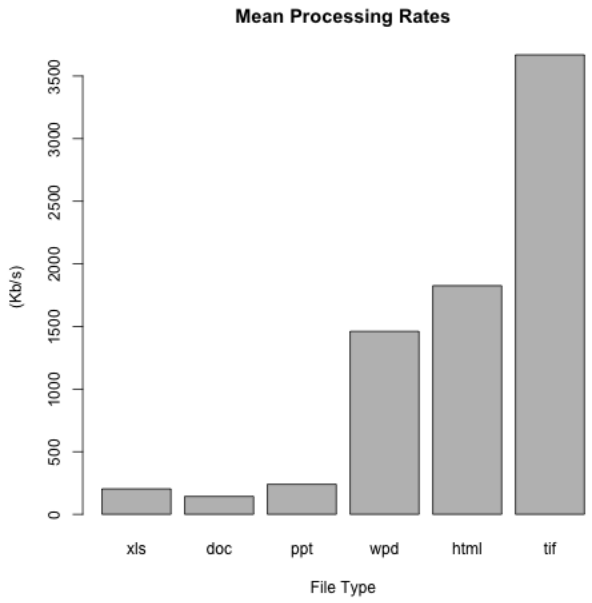


Figure 10.1.1: Mean Aggregate Processing Rates

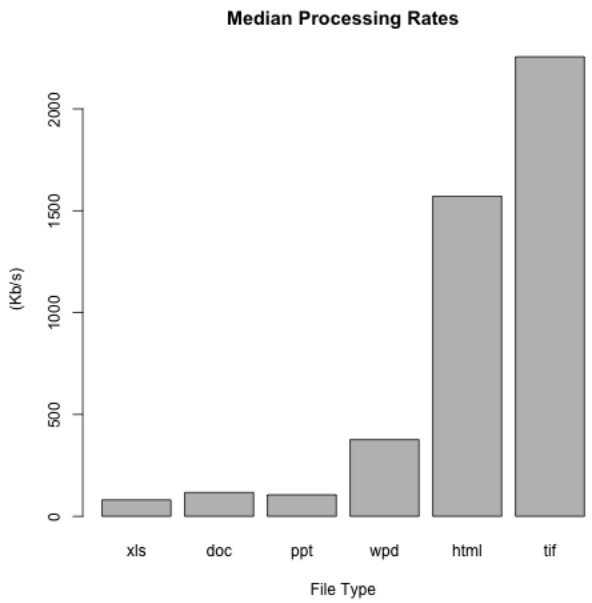


Figure 10.1.2: Median Aggregate Processing Rates

The aggregate processing rates for common file formats are included in figures 10.1.1 and 10.1.2. The Tif format is useful as a baseline because conversion to PDF is encapsulated rather than transformative in nature, thus the rate of processing is essentially theoretical maximum throughput. In common file types, the complexity of the file format significantly impacts the throughput; file formats like html are linear text with simplistic structure as opposed to Office Formats.

10.2 Comparison with Law Pre-Discovery

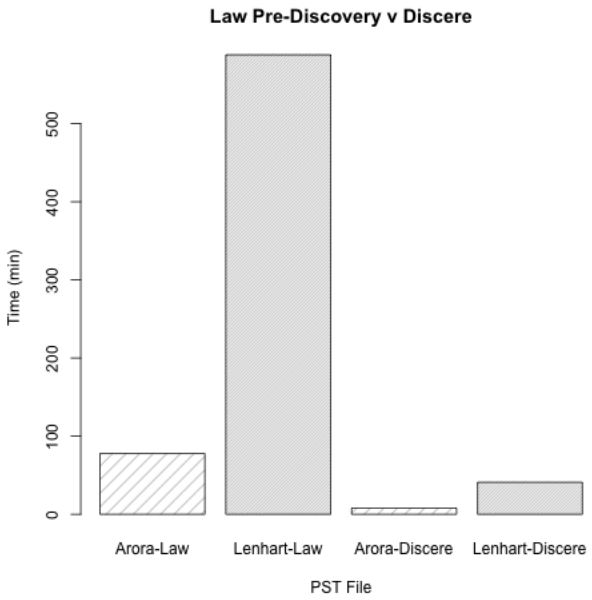


Figure 10.2.1: Comparison with Law Pre-Discovery

In the prior chapter, LAW was introduced as one of the most prominent tools in the industry for paginating ESI. Because the industry's existing state of the art is so inefficient, creating comparisons is very time consuming and resource intensive. My access to a LAW was limited to borrowing a single station with comparable resources to my own development system (i7 running at 3.4 Ghz, 16 GB RAM). I selected two files from the Enron corpus for comparison – zl_arora-h_000.pst and zl_lenhart-h_000.pst. The results from the comparison are presented in Figure 10.2.1.

Arora is approximately 67 MB in size while Lenhart is 1.52 GB. Lenhart has an abnormal number of multimedia files (audio and video) totaling 851 MB; when comparing LAW performance, it thus is closer to other PST files of approximately 700 MB. Discere outperforms LAW by a factor of between 9.7 and 14.3 respectively. Essentially, discounting communication overhead, 10-15 nodes of law would be required to match the performance of Discere in these instances.

10.3 Office Format

Office Format documents are, generally speaking, the set of documents commonly found in business use supported by the various office suites (Microsoft Office, Libre Office, Open Office, Pages, etc) originating in the Microsoft Office suite. The most common documents encountered in every day use are the .doc,

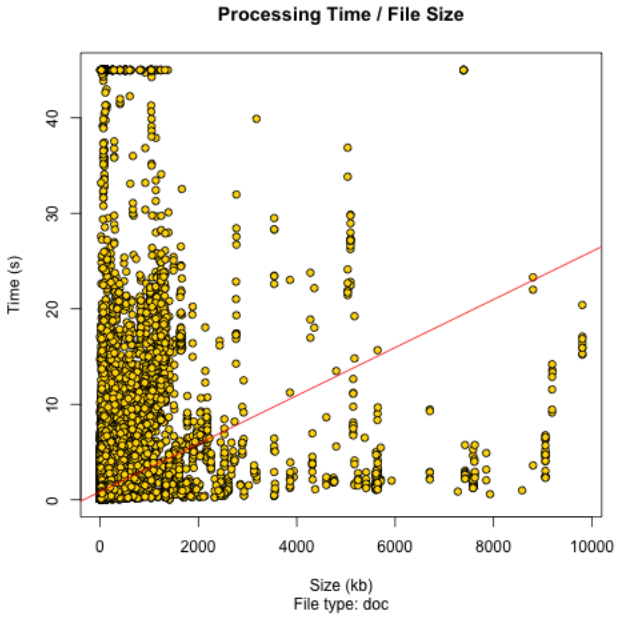


Figure 10.3.1: Doc performance

.xls, .ppt, and .wpd[1] formats. The sample data in the Enron data set predates the Office 2007 XML based formats (.docx, .xlsx, .pptx) so the analysis here does not include performance details for those specific format versions.

Word Document (.doc) performance (see figure 10.3.1) is highly varied, but with a general linear trend as file size increases. Though the file sizes in the data set are most dense between $[0, 2)$MB, the conversion time varies throughout the file sizes. The .doc file format is a Compound File Binary Format, meaning an individual .doc file may contain numerous file types. The presence of numerous file formats is to be expected given the versatility of Word Documents to include embedded visual media, macro programming, and other such data. The link between file size and conversion time, *ergo*, is correlative not causal. Conversion time is dependent on the content, and observations on general trends in a time *versus* size comparison hold true in as much as the sample is representative of general usage and composition of such formats in the wild.

Spread Sheet Document (.xls) performance (see figure 10.3.2) varies similarly to .doc with significant variance in conversion time based on the complexity of the document with some predictive value based on the spreadsheet size. While .xls is also a Compound File Binary Format , the variance is fairly dense throughout. Spreadsheets tend to have higher complexity not from included file types, but from complex formatting and/or a large number of individual worksheets within the document. For an example of spreadsheet complexity see figure 9.5.5 on page 121.

[1].wpd is a Word Perfect Document format. While .wpd has generally been supplanted by .doc, certain types of businesses still prefer the .wpd format for a variety of reasons. Law firms, in particular, still utilize this format heavily.

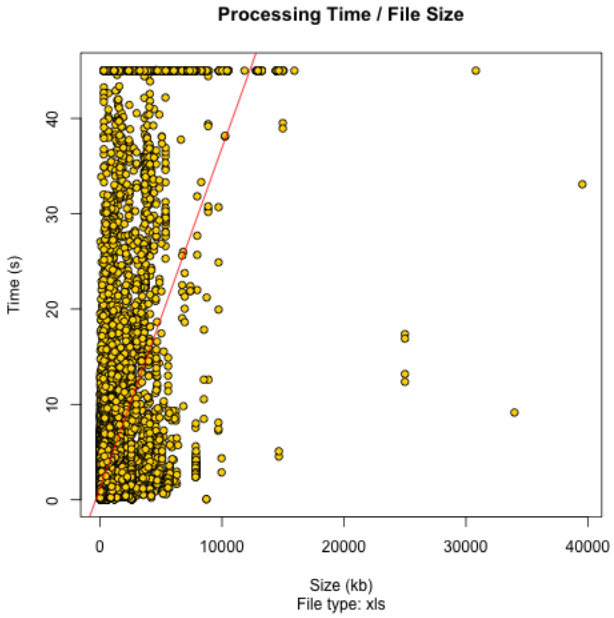


Figure 10.3.2: Xls performance

Word Perfect Document (.wpd) format has a more consistent processing time existing as a dense cluster of data points exhibiting near constant time (see figure 10.3.3 on the following page). As noted in 1 on the previous page, Word Perfect Documents have mostly been supplanted by .doc/.docx format, but still retain widespread use in legal contexts. In the case of the Enron dataset, this appears to be consistent. *Exempli gratia*, the email corpus for Kaminski[2][12] contains 91 .wpd files, consisting of text only, and primarily legal in nature. The consistent simplicity of the .wpd files present in the corpus can be attributed to the limited use and simplistic content these files are still in use for.

Power Point Presentation (.ppt) format, like the other Compound File Binary Format types, has a high variance in performance (see figure 10.3.4 on the following page). While .doc has a gradual regression slope, and .xls has a steep regression slope, .ppt files have a moderate regression slope. While the individual file will vary from the time projected by the regression line formula, the majority of files are less than 5 mb and processed under 20 seconds with the density of data points dropping outside those bounds significantly. This consistency makes the .ppt statistics more useful in estimating processing time *a priori* which will be relatively close, or proportionally close, to the *a posteriori* measured time. This *a priori* estimation is not as accurate when dealing with formats that have a high variance with high density in the varying range.

Rich Text Format (.rtf) files in this data set are fairly dense in both size and processing time (see figure 10.3.5 on page 138). The format does not support complex or compound components in the way that .doc, .xls, and .ppt do, and its performance is fairly consistent with very sparse outliers.

[2]zl_kaminski-v_000.pst-_003.pst, approximately 6.49 GB in total

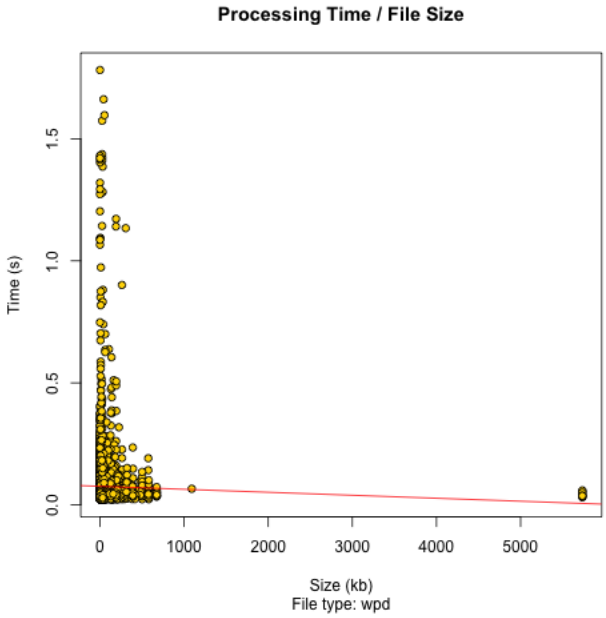


Figure 10.3.3: Wpd performance

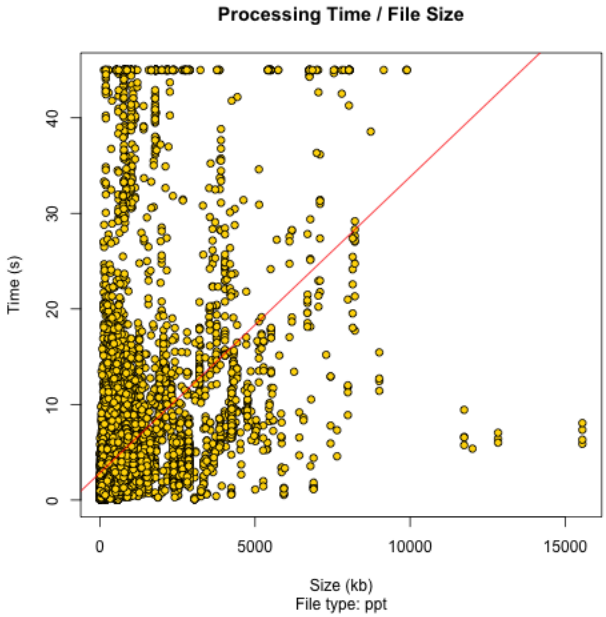


Figure 10.3.4: Ppt

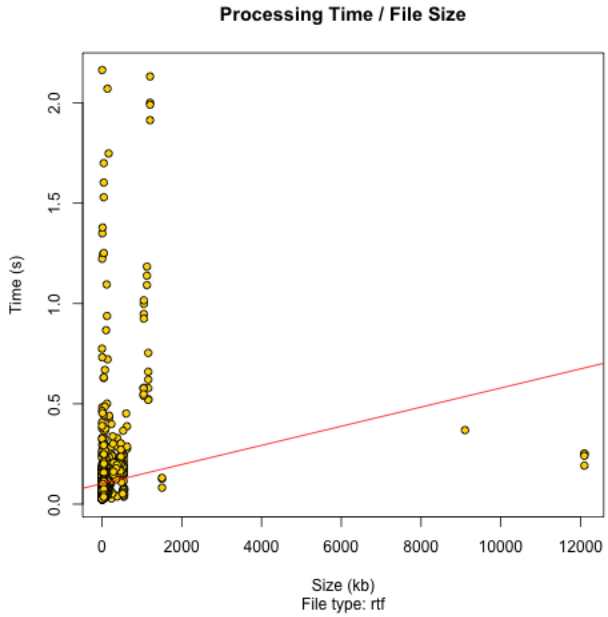


Figure 10.3.5: Rtf

10.4 Linear ASCII

There are primarily three types of ASCII files present in the data set which are human readable, and interpreted in a linear manner. These files include .htm and .html files which contain Hyper Text Markup Language ("HTML") formatting, and .txt text files which are pure human readable, non-formatted files. The .htm and .html (see figures 10.4.2 on the following page and 10.4.1 on the next page respectively) files are relatively small, quick to process, and few in number. They have a near constant regression line with little variance, and the data points are fairly small in number of the whole of the set compared to the office documents.

The .txt files (see figure 10.4.3 on page 140) are small in size with two main clusters in the sub 500k and the 1m range and a regression line similar to the .rtf regression line. There is some variance in the files, but not much, which is consistent with variances in performance based on the system running under load. Because the text files are formatless outside of line breaks, the variance exhibited by them while processed under load serves as a good control as to what variance is introduced into other metrics in this section from what the individual processing times would be if the individual conversions were run without parallel processing. Because a system such as Discere should always be run under load with processing done in parallel, metrics taken with the variance introduced by the I/O, CPU, and Memory/Paging bottlenecks are important for examining system performance in light of growing datasets.

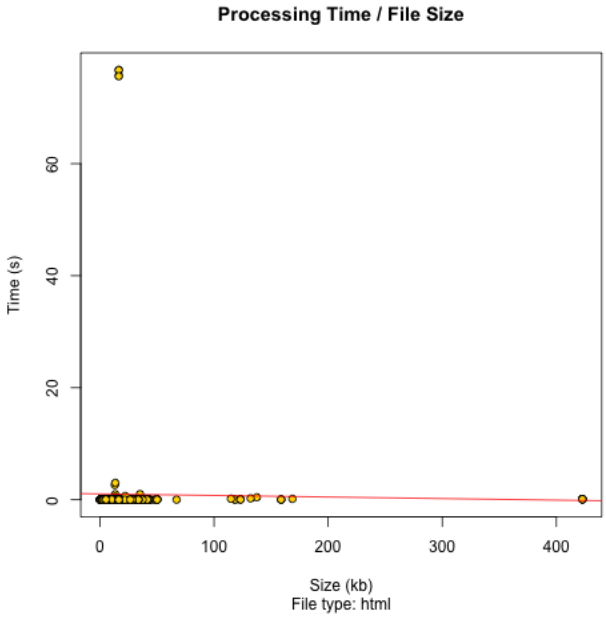


Figure 10.4.1: Html

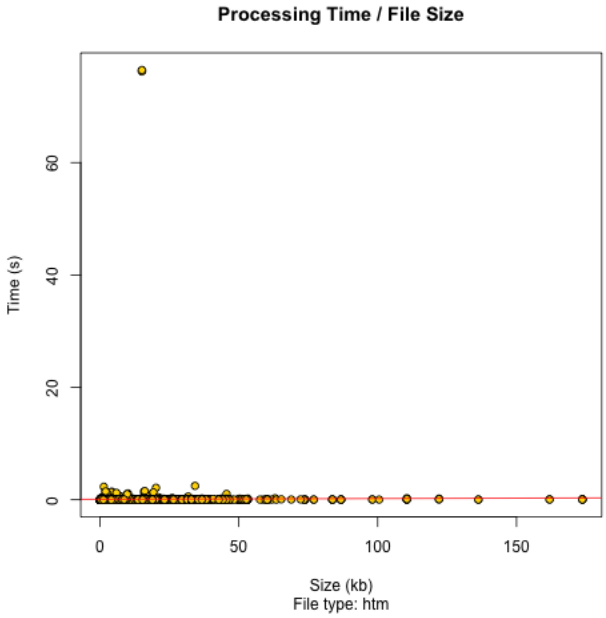


Figure 10.4.2: Htm

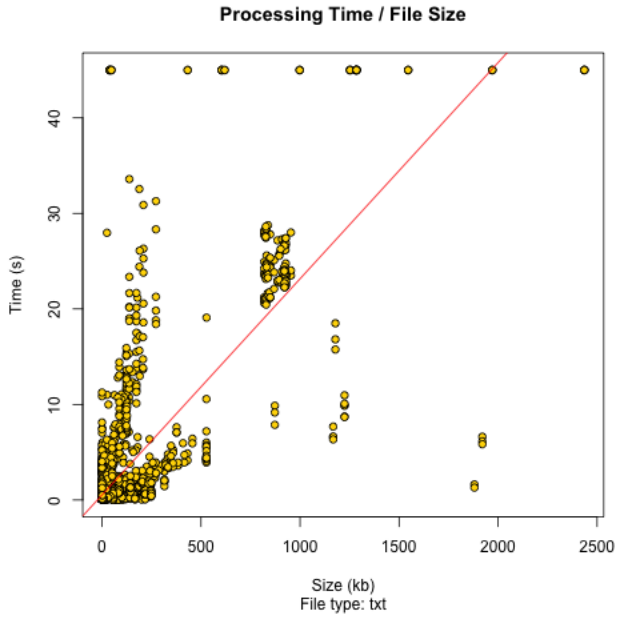


Figure 10.4.3: Txt

10.5 Image Format

Some image files are present, but are often irrelevant attachments especially where used as email signature blocks or corporate logos. Primarily, these images are .gif (see figure 10.5.1 on the next page), .jpg (see figure 10.5.3 on page 142, or .tif (see figure 10.5.2 on the next page). In general, .gif files are, as mentioned, more commonly found as email signatures, corporate logos or otherwise as part of an email background. Attachments of .jpg files typically are of photographs. Finally, .tif files are normally from a scan to email or other similar process.

The image files have relatively gradual increases in time as file size increases as evidenced by the gradual regression line slope. In context of other conversions, the image files are negligible contributors, both in time and quantity, to overall processing time, with jpg being the most prevalent in terms of quantity.

Other image file types such as .png and .bmp are too few in quantity over the entire corpus to be of statistical use even anecdotally. This may vary by organizational types such that organizations specializing in graphic media will have statistically relevant quantities of the images as well as larger images than a general business.

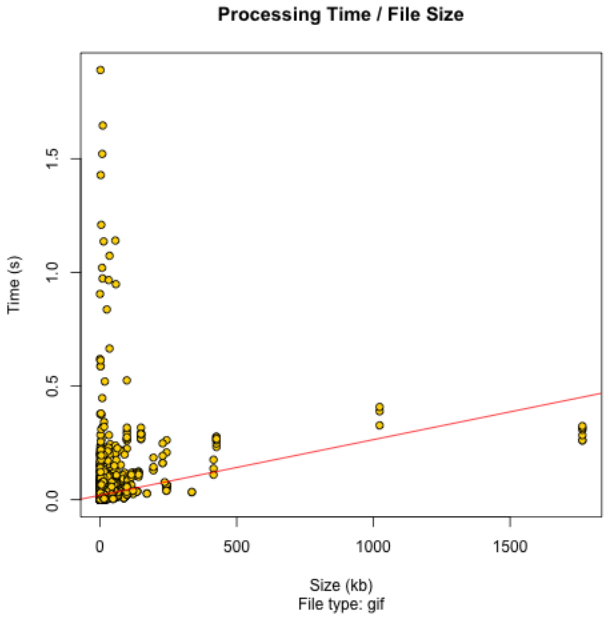


Figure 10.5.1: Gif

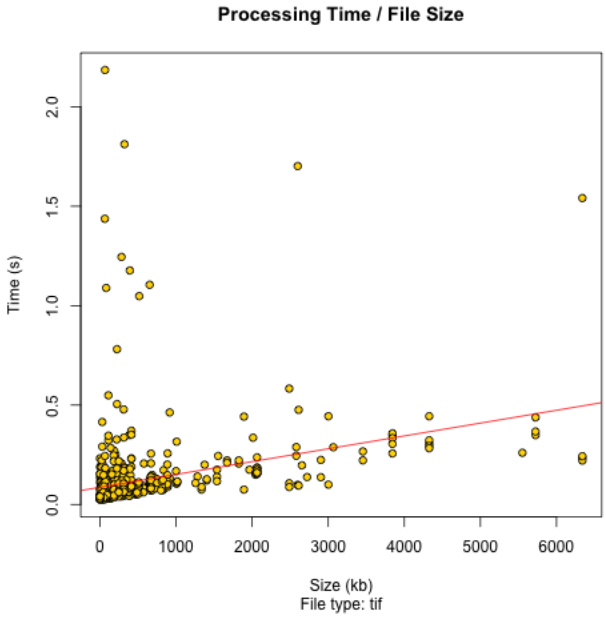


Figure 10.5.2: Tif

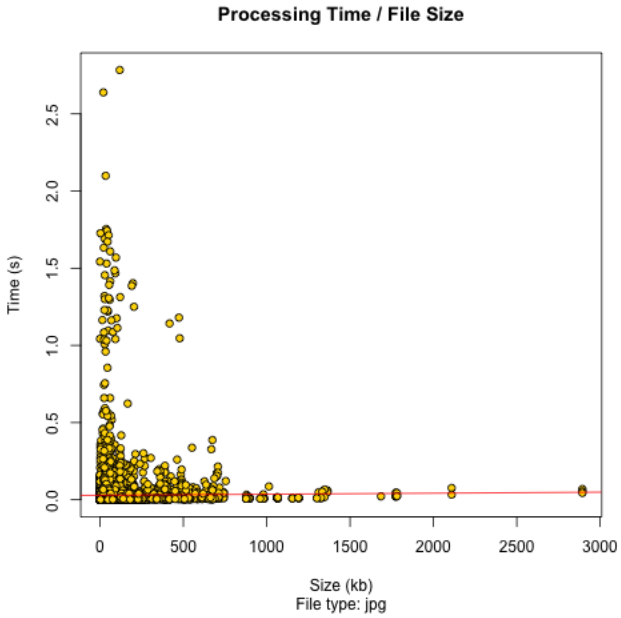


Figure 10.5.3: Jpg

10.6 Enron Whole-set

This chapter presents statistical results from the conversion process for specific common file types, as well as observations about the Enron data set itself in terms of pagination and file size comparisons. This information is important in trying to create *a priori* estimates which are relatively close to the *a posteriori* metrics in order to allow for accurate project management. Recall figure 9.2.2 on page 106 which details the processing node cost scalability of the leading commercial eDiscovery processing platform. For service providers, the cost scalability is necessary both to price their services accurately and to properly estimate the job requirements. For law firms, the cost scalability is necessary for determining whether they have the ability to conduct the conversion themselves with their existing in-house capacity or whether it must be sent to a specialized service provider to meet deadlines. In both cases, perfect *a posteriori* metrics are useless for making pre-processing decisions which have deadline implications for cases. Sufficiently accurate *a priori* estimates are required, but the question becomes whether statistical observations of known data set *a posteriori* results will allow for *a priori* predictions of future data sets.

This section examines the relationship between individual PST files in the Enron *corpus* and the properties of the post-processing rendering of said files. In all figures referenced in this section, each data point indicates the statistics for an individual PST file in the corpus. Each figure also plots a regression line based on the statistics of the set as a whole. Figures with data points densely clustered around the regression line indicate consistency with limited variance from predictable outcome; such high density / low variance is indicative of a metric which may fit the desire for *a priori* prediction of *a posteriori* metrics.

The number of pages a given data set will produce in paginated form is a basic question of high interest to law firms (who must review the documents) and service providers (who often bill per page). The nature of pagination makes estimates difficult to establish firm rules for. Page totals can be highly varied based on composition; a corpus with a high number of spreadsheets will often have an abnormally high page count while a corpus with large files not subject to pagination (such as video) will have an abnormally low page count. Despite variance in specific files, the Enron corpus shows a general trend (see figure 10.6.1 on the next page) for a linear increase. Variance becomes greater as the file size increases with the largest variance occurring for PST file sizes of 1GB and larger.

Total turnaround is an important factor in project management and capacity planning. Discere handles processing of emails separately from processing native files because it is more efficient to render paginated emails when doing the initial import rather than reading the data twice (once for import, and once for processing.) As such, the time metrics can be viewed separately (see figure 10.6.2 for import time including email processing, and figure 10.6.3 for native file conversion) or as a combined metric (see figure 10.6.4 on

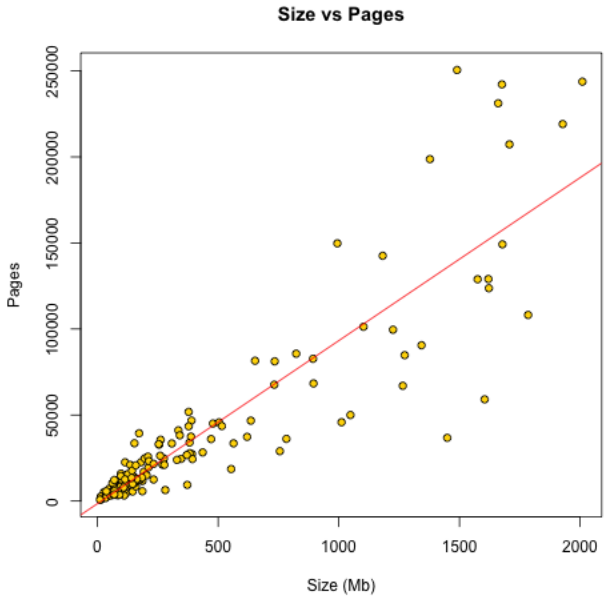


Figure 10.6.1: PST Size vs Page Total

page 146). Import time has a gradual slope as file size increases; this is primarily due to the nature of email rendering which is textual in nature and will experience the same processing increases as seen in .txt, .html, and .htm files (see figures 10.4.3, 10.4.1, and 10.4.2 respectively). Processing time has a steeper, but fairly linear, slope showing increases as file size increase consistent with the processing metrics observed in the main office file types detailed in section 10.3 on page 134. Given the overwhelming majority of file attachments found in the Enron corpus are .doc or .xls with .doc being the more numerous, it is not surprising that the processing time metrics follow the regression line of .doc files, and outliers with larger .xls content tend to be higher in processing time consistent with the steep .xls regression line slope. Given the larger proportion of time spent in processing native files into paginated formats, that the overall trend in total time matches up with the processing time metrics is expected.

The last metric comparison is the number of pages in the paginated rendering compared with the total time to process (see figure 10.6.5 on page 147). When compared with the results of PST size compared with page totals (see figure 10.6.1) and with the size versus total time (see figure 10.6.4 on page 146) a general trend can be observed. Both the page and the PST file size comparisons to total time produce similar regression lines with the former being *a posteriori* and the latter being *a priori*. We can observe that the *a priori* predictions using the regression line of the set tend to be slightly optimistic versus the actual *a posteriori* metrics based on total pages. This strongly implies initial size is predictive, but that size alone is not deterministic but rather is an abstraction of the effect the file type composition of the PST file has on

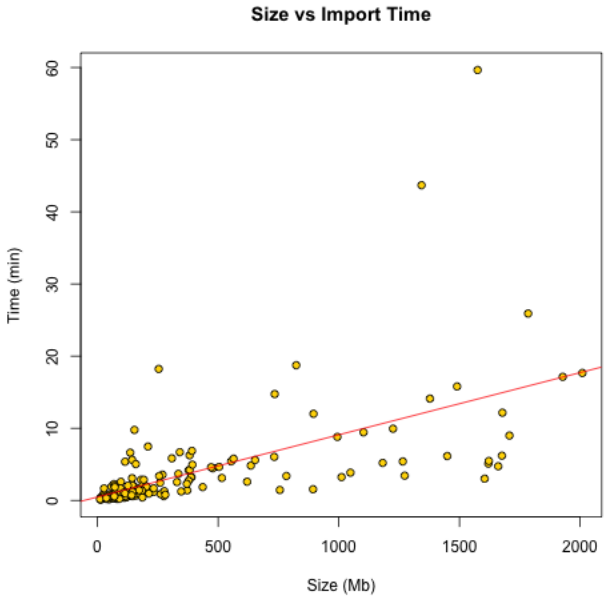


Figure 10.6.2: PST Size vs Import Time

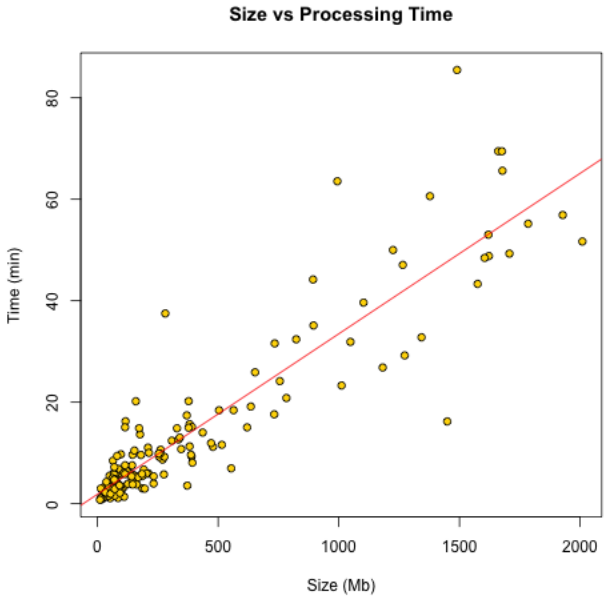


Figure 10.6.3: PST Size versus Processing Time

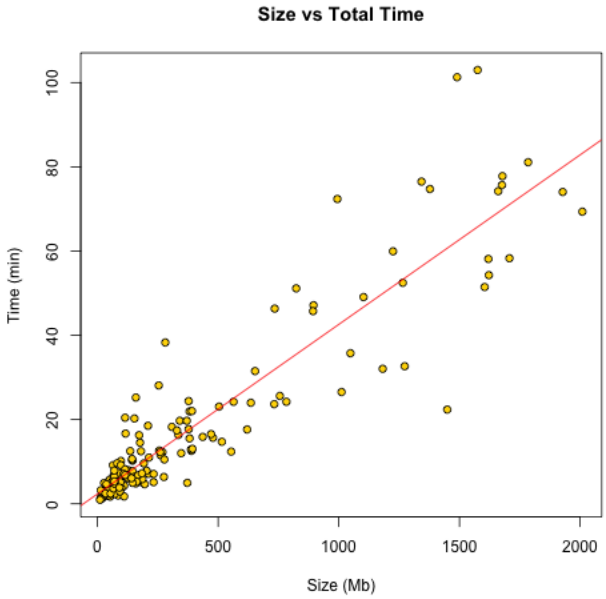


Figure 10.6.4: PST Size versus Total Time

total time and total pages after pagination. A general observation of pages per quantum (see figure 10.6.6) shows a varying range for the enron set most highly represented in the 50-150 pages/MB range.

10.7 TIFF *versus* PDF

As I discussed *supra* (see section 9.2.1 on page 106), the default format for pagination in commonly used eDiscovery tools is TIFF, and where PDF files are supported they are produced from TIFF rather than from textual construction. Discere takes the opposite approach by creating PDF files directly from text using either manual construction for textual sources (such as text files, or email) or by using libraries, applications associated with the file type, or other measures with direct PDF conversion capabilities for the file type. For this section, the PDF result of converting the Enron was then converted to TIFF to examine the impact TIFF has for scalability.[20, 16]

When we look at the size comparisons between PST file size and the size of the converted PDF result (see figure 10.7.1 on page 148) we find the PDF result is equivalent or slightly smaller than the PST native file. This occurs because PDF conversion directly from text is extremely efficient from a storage perspective, and because some of the PST file size is occupied by unsupported or non-paginatable file types (thus being represented by a placeholder file) which include such things as video files (which, in some cases, take up significant storage and result in outliers trending to the smaller size). There are also outliers which are larger than the PST file size especially where there is a disproportionate inclusion of spreadsheet files or

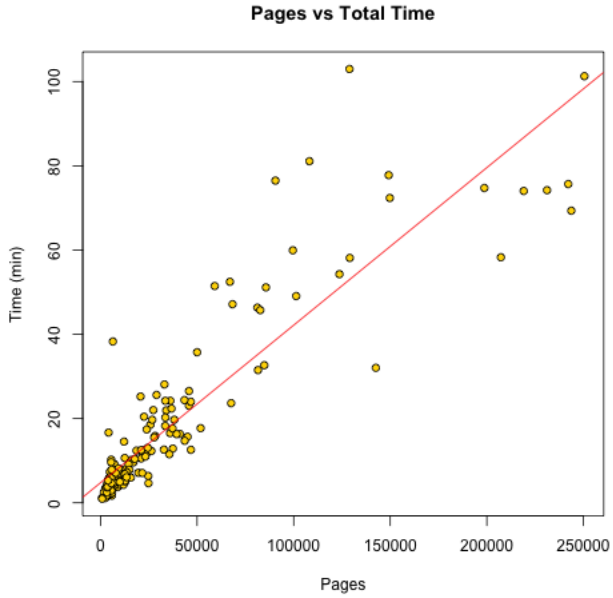


Figure 10.6.5: Pages versus Total Time

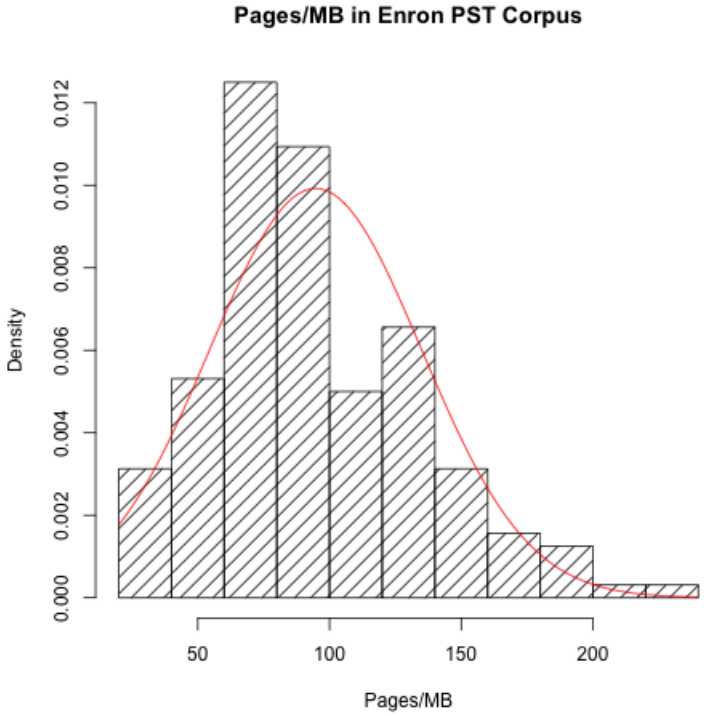


Figure 10.6.6: Pages per MB Distribution

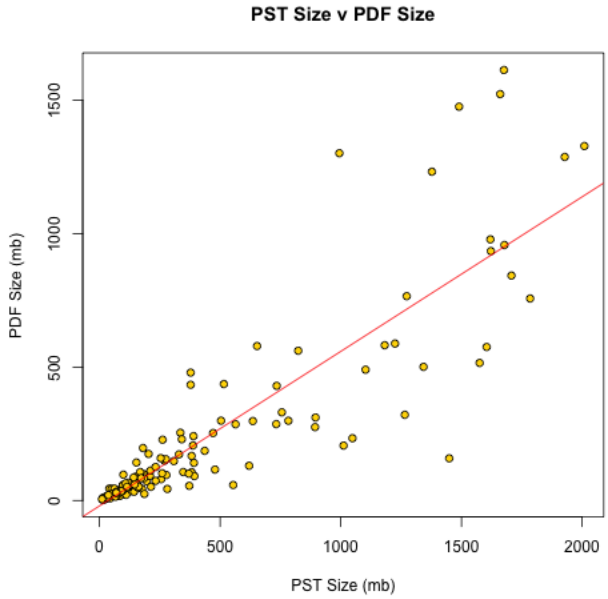


Figure 10.7.1: PST vs PDF

other similar formats that produce significant page counts. In either case, the differences in size between native and PDF is in the same relative range.

Comparing the PST size to a TIFF rendering of the PDF set (see figure 10.7.2 on the next page) shows a substantial increase in size for the *corpus*. Typically, the TIFF corpus is at least an order of magnitude larger than native PST itself so a 1GB PST file will produce something in the neighborhood of 10GB of TIFF files on average. This significant increase in size makes TIFF untenable as a format for use in the future as the increased storage requirements create unnecessary I/O bottlenecks, the increased CPU requirements needlessly increases processing time and hardware requirements, and there is an overall difficulty created in transporting the data.

In comparison to the equivalent data in PDF format (see figure 10.7.3 on the following page), the use of TIFF files presents an unnecessary burden at all stages of the discovery process.

10.8 Conclusions

Performance for any conversion tool for any data set is going to be dependent on the file type composition of the data set, the performance of that tool for the given file types, and the complexity of the specific file. Discere performs well for all categories of supported file types. Further, by converting files directly to PDF and preserving textual information Discere eliminates the need for later OCR and reduces the storage footprint of the processed result set by, on average, an order of magnitude. Processing time itself is between

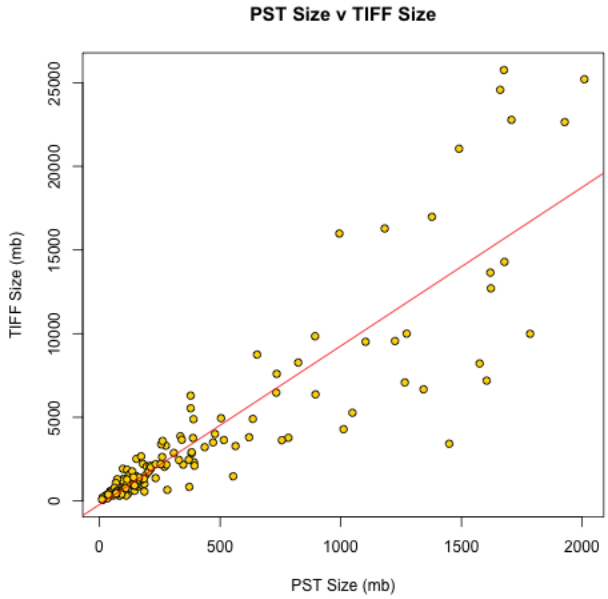


Figure 10.7.2: PST vs TIFF

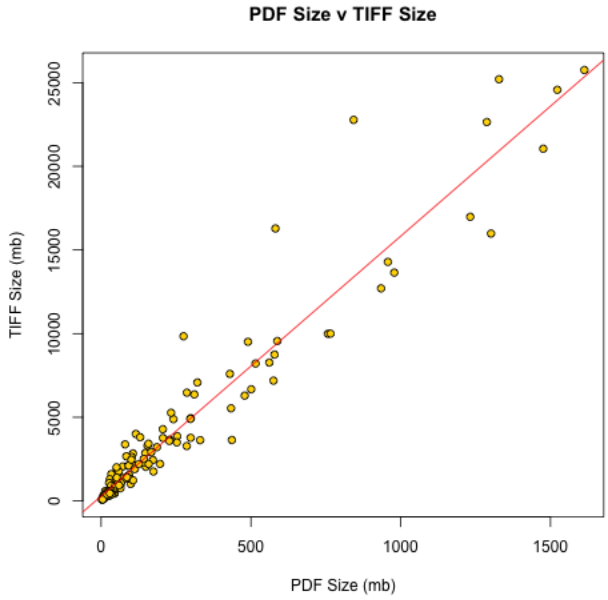


Figure 10.7.3: PDF vs TIFF

one and two orders of magnitude from what current widely used commercial tools are capable of today.

Estimating the time a particular set of files will take to convert *a priori* is a difficult endeavor as demonstrated by the results in this chapter. Because a number of file types have a large variance in the same size range, composition percentages of the data set will vary the accuracy of predictions, while other file types have a fairly predictable time/size increase slope. But, the larger the data set being estimated the more likely the effect of outliers will be minimized, and the more accurate the prediction. Because, in most instances, predictability is needed primarily in larger datasets, the variability of certain file sizes will not have as large of an impact and the predictions based on individual file type / size / time observations will allow sufficiently accurate estimates for project management and scheduling purposes.

Chapter 11

Scalability

Up until this point, my presentation of Discere as a solution to the problems previously identified has centered around a single user paradigm. I chose the single user view as the one most likely to find my prototype useful for real world adoption as those use cases are the ones most likely, in my experience, to not have any tool at all. That being said, my treatment of the problems in this realm have not been to the exclusion of scalability concerns in reaching a multi-user use case with significantly larger data sets. This chapter identifies further existing problems in how the industry currently treats electronic data for legacy reasons left over from a physical paper work flow, how Discere's underlying indexing and search systems are easily extended for distributed and client-server based architectures, and a roadmap of future work for overall extension of Discere into a larger framework.

11.1 eBates

A 'Bates' number gets its name from a mechanical stamp which increments its displayed digits sequentially as pages in a document are stamped. The numbering system was created to provide a way for attorneys to refer to specific pages and documents as well as ensure pages do not go missing. The system continues to be used though modern usage has replaced the physical stamp with a digital process of labeling document page images. In a scanning based system in which physical documents are digitally images, the stamping model continues to work, but electronic documents generally have no pre-existing paginated structure until rendered into a printable format. This produces difficulties in determining, *a priori*, what beginning number a given document will start with *a posteriori.* These difficulties are experienced where an electronic *corpus* is to be rendered and numbered in one step, where a given rendering must be replaced in an already rendered and numbered *corpus*, and where a given electronic sources lacks clear pagination or where it may have multiple possible paginations.

When a set of electronic documents $\{D_0 ... D_n\}$ are to be rendered (to PDF or Tiff) and numbered

(sequential numbering displayed on each page) the process can be completed in one step in a single threaded / single node process without diverging from the classic numbering model, but increasing the number of nodes or multi-threading the process makes it incompatible with classic numbering and requires a two-stage process. Document D_1 will have a first page whose number is one greater than the last page in D_0, likewise Document D_n will have its first page numbered one greater than the last page in D_{n-1}. In a single thread / single node process, D_0 will be complete before D_1 is started, such that D_{n-1} will always be complete before D_n begins. Thus, the process will always be aware of what the page number for the currently rendered page should be before moving to the next. If, however, more than one thread or nodes are being used in the process, the Document set $\{D\}$ will necessarily be partitioned into $\{D_0...D_i\}$ and $\{D_{i+1}...D_n\}$ in some manner. If D_0 is processed by one thread, we know its first page will be 1 and each subsequent page will be a single increment as we render them. On the other hand, we will not know the first page of D_1 or any other D until D_0 completes because electronic files do not, generally, have a known-page *a priori* page count. Thus, to reap the benefits of multi-threading or multiple nodes, the process must be completed in two stages with Stage 1 rendering paginated versions of the electronic files, and Stage 2 applying sequential page numbers.

Now, let us assume $\{D\}$ has been rendered into a paginated form $\{P\}$ such that for any D_i there is a corresponding P_i. If it is found D_x has a defective pagination before $\{P\}$ is numbered, switching out $P_{x,old}$ for $P_{x,new}$ creates minimal impact on the process. If $\{P\}$ is already numbered, and $P_{x,new}$ has more or less pages than $P_{x,old}$ then all $\{P_{x+1}...P_n\}$ will have to be re-numbered. In cases where the paginated format is primarily image based rather than textual, this requires significant I/O bandwidth to re-write from the original un-numbered copy or re-process the native files to fix the numbering error. Likewise, no D may be removed save D_n without disturbing the page numbering either.

Unlike physical documents which have a clear pagination, electronic documents do not necessarily have a pagination at all which corresponds to the document's nature. An example of this type of unpaginated electronic document is the database. A database contains data which may be arbitrarily queried into a report or other paginateable rendering, but that rendering is not the totality of the database itself. Some information in a database is not intended for human consumption, but for machine consumption and may involve access logs, user tables, and the like. In this last case of amorphous file types presents a real problem for a legacy approach based on the physical paper document paradigm.

The incongruity between physical documents and electronic data confounds traditional numbering in a way that seems minimally problematic to the laymen in their human consumption of these number, but significantly problematic to distributed processing and scalability of the electronic discovery process. Worse, the rigid numbering system behaves like an array with the innate inefficiencies of insertion. These problems are not insurmountable, but have presently not been addressed by existing approaches.

11.2 Distributed Processing Compatible Electronic Numbering

To solve the numbering problem introduced by an inability to determine page count *a priori,* the underlying purpose behind the numbering system must be examined. To this end, we must ask why sequential numbering is useful, in what ways it is useful, and if system compatible with distributed processing can be devised which retains its useful features. A unique sequential number allows three benefits to a human interpreter - (1) it allows a specific document or page to be referred to easily, (2) it allows for fast location of that specific document in an ordered *corpus*, and (3) it allows detection of missing pages.

Some aspects of the sequential numbering have become or will soon become unnecessary. With even paginated versions of electronic data widely stored in electronic form rather than being produced in physical format, the detection of missing pages due to being torn out, dropped, or otherwise mislaid is no longer a necessary feature. Likewise, an electronic indices of documents divorces the human comprehensible identifier from the mechanism used to search or otherwise locate the document itself - *id est* the identifier APPLE is just as functional as the identifier 12345 in identifying and recalling the document if the identifier is part of an electronic search system rather than a manual flip through a set of ordered pages. There is some merit, however, in knowing that a given document is sequentially close or far from another given document if the order of documents has some meaning such as source, custodian, or type.

We know how many documents we have, but we do not know how many pages they will render into. In a page-sequential numbering system, D_1's numbering is dependent on the final page number of D_0 *a posteriori,* but we know *a priori* the sequential document number of following documents. The answer then, it seems, is to separate the numbering schema in such a way that the first page of D_n is independent of D_{n-1}. This can be accomplished by introducing two dimensional representation. In section 11.1 on page 151, I noted the sequential numbering system is similar to an array base data structure in that it has an innate difficulty with insertion and changes in numbering. In a similar analogy, we might consider an array of arrays in that the given element is the document D_i and the given pages for the document are separately identified $D_{i,1\ldots n}$ thus the first page of document D_i is 1, as is the first page of every D. Instead of a single sequential numbering of pages, the numbering can be segmented into a sequential numbering of documents, and within each document a sequential ordering of pages forming $D.P$ such that the 5th page of the 9th document might be 9.5 rather than 48 or some such. In such a way, the documents may be distributed across nodes and threads without concern for the order in which they are processed and numbered.

Electronic data which is not paginateable, has multiple possible paginations, or is only partially paginateable can be better handled in this system. First, by separating the page numbering from the document numbering we retain the ability to identify specific document files and to refer to specific data points within

those files by byte offset or range without relying on page renderings especially where certain data attributes might not be rendered as in the human viewable rendering. Second, where multiple renderings are possible we may specify an optional rendering component such that $D : R.P$ can refer to a page within the specific rendering of a document.

By separating the traditional sequential page numbering into discrete components integrated hierarchically, we extend a previously problematic legacy numbering system into a more precise system which allows us to differentiate between the original file with its human readable and non-human readable component data, and to handle file formats which do not have a single easily identifiable way to render a pagination.

Even in file types which seem to have easy paths to pagination like, for example, a Word document, there are potential uses for alternate renderings especially where digital forensic analysis is engaged. Consider that Word documents may contain fragments of or complete prior versions of the current presentation of the documents, and that data may exist within the document file. We can thus envision a use-case for differentiating between $D_i : R_{default}.1 - n$ (Pages 1-n, of the default rendering of document i) versus $D_i : R_{prior}.1 - n$ (Pages 1-n, of a prior version fragment of document i). With each component being independent of the other components, the numbering system would not need to be distributed nor would other documents be affected by discovering prior versions of the 'document' contained within the electronic file.

By recognizing the paginated rendering of an electronic file is different from the file itself, and accounting for this divergence in the numbering system, we can adapt such a system to be both more useful in describing documents as well as more efficient in allowing for better methods of distributed processing. The specific examples here are illustrative, but not exhaustive of how these components could be represented; certainly using electronic review tools allows for some aspects to be hidden - calling up a document number might show a default rendering, but graphically indicate other renderings are available, and might further still provide for access to the native electronic file. The numbering system can further break down the document number to have a prefix indicating the source or custodian of the data, the case, or even which party produced it. An extensible hierarchical representation thus allows for consistent representation globally rather than in a limited myopic view of a single case or party.

11.3 Client-Server Architecture

At its inception, Discere was designed to address two problems in the industry. First, existing approaches are very inefficient both in time complexity and space complexity. Second, existing solutions are often prohibitively expensive for the small firm or solo legal practitioner. Solutions to the first problem are

independent of the target market, but solutions to the second problem necessarily will involve a different architecture (in the previously presented examples, a single-user paradigm) than is necessary to support larger (potentially decentralized or geographically separate) user base, significantly larger data sets, and complex long-term cases. To address the second problem in a way which allows for incrementally increasing the technical complexity of the architecture while still allowing for the same simplistic solution for the small or solo user base, selecting appropriate technology to facilitate the indexing and review aspect was vital.

What differentiates a small firm or solo practitioner from a medium or large firm? In much the same way small businesses differ from medium and large businesses, small and solo firms have significantly less access to technical personnel and expertise than larger organizations. The concept of 'IT' has become a catch all for any technical employee and encompasses front line help desk personnel and low level technicians that keep desktops and laptops running, set up simple wireless networks, and install varieties of common software applications as well as highly skilled and experienced system and network administrators who are experienced with managing complex network topologies, server clusters, and the like. Where a small organization might have a file server, and an email server, a larger organization has numerous email servers with different roles[1]. The smaller organization will not have the expertise or resources to set up and manage an elaborate distributed, load balanced, or otherwise high performance system, nor do they have the need. The larger organization, on the other hand, cannot survive with a single point of failure and, with high volume, can easily overwhelm a single physical server.

The solution lies in Discere's use of the Lucene project as the core indexing and search system.[42] In 2010 the Apache Lucene project merged with the Apache Solr project[4] providing for a seamless way to apply the same methods of indexing and review used in a single user version using Lucene as the local backend, to client-server architecture using Solr as the backend. One of the overriding ideas behind Discere was that there are existing open-source projects which provide the capabilities needed in the complex undertaking eDiscovery processing and review tools represent, and by carefully selecting the right tools for each aspect the continued advancements of those projects could be harnessed without duplicating effort to rewrite support for capabilities that already exists. As the Solr project grew and matured eventually to merge with Lucene, Discere gained the ability to extend itself into a client-server architecture with minimal effort simply by taking advantage of the distributed, load balancing, and other clustering features the Solr project introduced.

By using Lucene for a single user paradigm, Discere can be deployed to individual workstations without

[1]Modern Microsoft Exchange setups in larger organizations have discrete servers handling internet facing reception or sending of email (Edge Transport Server), storage of the actual user mailboxes (Mailbox Server), email access be it the Outlook Web App, Exchange ActiveSync, traditional POP3 or IMAP4 (Client Access Server), and additional roles tying the architecture together (Hub Transport Server) or providing integration with PBX or Fax systems (Unified Messaging Server). Contrast this with a smaller organization which may localize all these functions on one system (if using exchange, or pre-integrated using Small Business Server) or use a different solution centralized on one server.

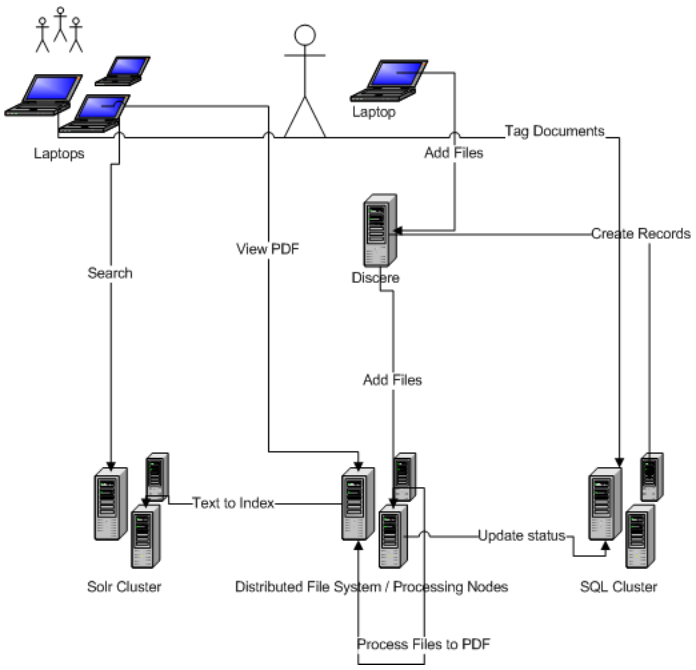


Figure 11.3.1: Distributed Architecture Example

any thought about servers, networks, or the like. At the same time, by using the Solr project the same solution which serves a single user can be extended by changing the target to a Solr system with the only changes being how and where the files are stored (locally vs remotely) while the extraneous details of how many server shared, replication, load balancing and the like are handled by the Solr project itself abstractly from the client. While this functionality was not necessary to solve the problems I set out to tackle, the ability to easily scale up the architecture makes the solution universally applicable rather than confining it the smaller use case I envisioned originally.

11.4 Future Work

Discere as a single user solution has gone beyond the proof of concept I originally intended, and is a fully functional prototype ready for real users. There are still further applications I see for it which I intend to pursue in the future to expand its capabilities as well as address certain use cases I still think are missing in both the industry and open source markets.

First, there is a gap between open source forensics tools and electronic discovery tools. The more prevalent open source forensic tool is The Sleuth Kit ("TSK") and its Autopsy Forensic Browser. Many open source forensic tools in someway utilize TSK, but those tools have rudimentary export functions for making meaningful exports of an investigation into a discovery review platform. TSK is available on many

platforms, and Autopsy's current development branch is preparing to deploy version 3 as a Java based application. Autopsy 3 is already utilizing SOLR/Lucene for its indexing making the incorporation of an export function there or an import function in Discere entirely viable in the future to complete the tool chain from ingestion of raw forensic image data to processing/review of documents by attorneys, to production to opposing counsel in response to discovery obligations or requests.

Second, though the single user version of Discere is in the working prototype stage, Solr is still very much a proof of concept. Future work will see the Solr backend developed as a production ready option along with implementation and best practice guides for administrators. There are also concerns in a Solr based model not present in a single user version in tracking remote client information (IP, user, time, etc) for auditing purposes.

Third, distributed processing of large data sets is something that, ideally, would take place in a cluster rather than involving desktop or workstations especially in a system using a Solr backend. To that point, after a Solr solution is production ready the next logical step would be to incorporate a distributed conversion system that supports job submission rather than relying on a GUI control system. Potentially a Hadoop based system may be useful for this especially in light of Solr development which has begun making Hadoop based indexing feasible.

The future work road map I have laid out sets a course to take the solution to the existing problems and expand it into a large scale platform suitable for medium to large organizations and user bases. Much of what makes this road map feasible is the continued advances in the underlying components I have selected for solving the problems I originally identified. What originally would have been infeasible has become feasible because of the continued efforts of the Open Source communities, most notably the Apache Software Foundation, in integrating related projects in such a way as to make them suitable as a larger framework.

In conclusion...

The catalyst for changing the Federal Rules of Civil Procedure in 2006 was, undisputedly, the failures in the Zubulake cases. My analysis of the Zubulake and Metropolitan Opera cases highlights the danger inherent in mishandling ESI, the difficulty in detecting failures, and the severe consequences failures can have. Successive cases interpreting the new rules have demonstrated a range of possible sanctions for discovery failures and numerous different factors used to evaluate discovery obligations.

My analysis of case law clearly shows the courts, lawyers, and parties are still struggling with the extent of discovery obligations as well as the ephemeral nature of Electronically Stored Information. Handling electronic data is counterintuitive if it is handled with the same methodologies used for physical documents; the physical analogy simply cannot model the underlying concerns sufficiency regarding access to the data, awareness of data sources, and the ability to search for responsive documents.

The case studies I presented in Part II illustrate the dangers of failure in these three areas. As we saw, assumptions about an underlying system can be problematic, but such incorrect assumptions are no excuse for failing to meet discovery obligations. The real danger, however, is the severe difficulty a party currently has in demonstrating the opposing side failed to meet their obligations. This was the problem faced in Zubulake, and the problem which continues into the future.

The corollary difficulty faced by the producing party is adequately protecting a client's interests by reviewing and excluding privileged information. As data sets grow, the feasibility of this continuing as a purely manual process is untenable. The same expertise needed to properly locate responsive documents must also be applied to excluding privileged ones from the responsive sets. Cost, however, is often the determining factor in how such problems will be approached.

The criteria used to balance the reasonableness of discovery practices often looks to the burden that would be imposed. Existing approaches are inefficient, costly, and slow; these factors directly impact the burden the courts will consider in determining reasonableness. In effect, the worse the state of the art, the more limited the scope of discovery, and the greater the potential for justice being denied. The industry currently employes closed source commercial, proprietary, or rebranded closed-source commercial tools. These tools are costly, scale poorly, and are not transparent in how they function.

High cost is exacerbated by these tools' inefficiency and their use trends toward limiting discovery unnecessarily. Worse, smaller firms often do not use any tools at all, but rather perform labor intensive and error prone manual reviews ad-hoc. Disparate adoption, closed source products with opaque practices, and a lack of standardization create unnecessary difficulties in proving sufficiency, and detecting insufficiency in discovery practices.

Essentially, the current state of the industry creates a black box where collection, processing, review, and production stages of discovery are divorced from each other. The lack of transparency in how these tools function combined with the nature of the discovery process relying on the honor system makes it very difficult to detect discovery abuses. As demonstrated in my case studies, often a discovery failure may not even be cognizable by the legal team responsible for collection because how the underlying search system or query syntax works is unclear to them. Even the nature of the business practices that populate a search system play a significant factor into how a search must be designed. What the underlying system contains and how it searches are vital in evaluating the sufficiency of searches in context of whether those searches meet discovery obligations and fully answer production requests.

When I set out to build Discere, my hope was to demonstrate the viability of open source alternatives to the extremely costly proprietary tools currently in use and to show the benefits of integrating processing, review, and production into a single cross-platform application. By identifying systemic flaws in the approach used in proprietary tools, and selecting libraries which allowed low level manipulation of the pertinent file formats, Discere well exceeded these modest original goals.

Discere, running on a single system, outperforms cluster based proprietary solutions by between one and two orders of magnitude. The cost savings of using open source tools, and the self contained nature of the tool make it suitable for small firms that currently use nothing and exceed the capabilities of large firms and specialty litigation support vendors. The transparency of the search system, its query syntax, and the index backend allows searches to be evaluated against known standards rather than educated guesses based on observed search behavior or results; transparency is vital to defending sufficiency and detecting failures.

Superior performance, integrating all phases of discovery, and transparent functionality create the potential for wide adoption. Because there is a significant portion of the industry for whom existing solutions are not feasible, it also has great utility as an equalizer between large firms with significant resources and smaller firms which do not have the same resources.

My contribution to the field herein is, therefore, multifold. First, I have identified the impact rules, trends, and case law have which governs our participation in the discovery process both aiding our clients to acquire the data they are obliged to produce and to detect deficiencies in productions by their opposing counsel. Second, I have demonstrated the concerns about sufficiency are real and have significant potential

to facilitate or deny the cause of justice. Third, I have distilled and articulated many of the issues involved in precise and, where possible, mathematical ways to facilitate better explanations to the courts of what is or is not sufficient, is or is not burdensome, or is or is not feasible. Finally, I have introduced new tools to materially reduce the burden Electronically Stored Information heretofore represented. In so doing, I have tackled an emerging portion of our field which has been left largely unexplored.

QUOD ERAT DEMONSTRANDUM

Index rerum notabilium

Bibliography

[1] American friends of yeshivat ohr yerushalayim inc v US.pdf.

[2] Apache pdf box.

[3] Apache poi http://poi.apache.org/.

[4] Apache solr - solr.apache.org.

[5] Apache tika - a content analysis toolkit - http://tika.apache.org/.

[6] Arista records LLC v. usenet.com, inc., 608 F.Supp.2d 409 (2009).

[7] Carrot 2 search results clustering engine.

[8] Convolve inc v compaq computer corp.pdf.

[9] The DAT roadmap. http://www.dat-mgm.com/DAT%20roadmap/.

[10] De espana v american bureau of shipping.pdf.

[11] Ehrenhaus v. reynolds, 965 f.2d 916, 921 (10th cir. 1992).

[12] Enron dataset.

[13] *Federal Rules of Civil Procedure*.

[14] *Federal Rules of Evidence*, 2010 edition.

[15] Felman production, inc. v. industrial risk insurers, slip copy (2010).

[16] Ghostscript 9.05.

[17] Harkabi v. SanDisk corp., 275 F.R.D. 414 (2010).

[18] Hickman v. taylor, 329 U.S. 495 (1947).

[19] Ice pdf viewer - http://icepdf-viewer.icesoft.org/icepdf-viewer/icepdfviewer.iface.

[20] Image magick 6.7.7-8.

[21] itext pdf - http://itextpdf.com/.

[22] Java-libpst - http://code.google.com/p/java-libpst/.

[23] Jodconverter - http://code.google.com/p/jodconverter/.

[24] Lee v. max intern., LLC (2010) WL 2680429.

[25] Lee v. max intern., LLC (2010) WL 2680439.

[26] Lee v. max intern., LLC, 638 f.3d 1318 (2011).

[27] Mancia v. mayflower textile servs. co., (2009) WL 2252151.

[28] Mancia v. mayflower textile servs. co., 253 F.R.D. 354 (2008).

[29] Metropolitan opera ass'n, inc. v. local 100, hotel employees and restaurant employees international union, et al., 212 F.R.D. 178 (2003).

[30] Rhoads indus., inc. v. bldg. materials corp. of am., 254 F.R.D. 216 order clarified, 254 F.R.D. 238 (E.D. pa. 2008).

[31] Richard green (Fine paintings) v McClendon.pdf.

[32] S.E.C. v. collins & aikman corp., 256 F.R.D. 403 (S.D.N.Y. 2009).

[33] Standard ECMA-320. http://www.ecma-international.org/publications/standards/Ecma-320.htm.

[34] Ultrium - LTO technology - ultrium GenerationsLTO. http://www.ultrium.com/technology/generations.html.

[35] Victor stanley, inc. v. creative pipe, inc., 250 F.R.D. 251 (D. md. 2008).

[36] Young again products, inc. v. acord, 09-1481, 2011 wl 6450843 (4th cir. dec. 23, 2011).

[37] Zubulake v. UBS warburg LLC, 216 F.R.D. 280 (2003).

[38] Zubulake v. UBS warburg LLC, 217 F.R.D. 309 (2003).

[39] Zubulake v. UBS warburg LLC, 220 F.R.D. 212 (2003). Zubulake IV.

[40] Zubulake v. UBS warburg LLC, 229 F.R.D. 422 (2004).

[41] 04 2012.

[42] Apache Software Foundation. Apache lucene.

[43] E.G. Fayen F.W. Lancaster. *Information Retrieval On-Line*. Melville Publishing Co., 1973.

[44] David Glovin. UBS must pay Ex-Saleswoman $29.3 mln in sex bias case (Update6) - bloomberg.

[45] John Markoff Katie Hafner. *CYBERPUNK: Outlaws and Hackers on the Computer Frontier, Revised*. Simon and Schuster, 1995.

[46] Charles Kozierok. PCGuide - ref - "Just a bunch of disks". http://www.pcguide.com/ref/hdd/perf/raid/levels/jbod.htm, April 2001.

[47] Brian Roux. Changes in approach for scalability in digital forensic analysis emphasizing distributed, collaborative, and automated techniques. Speaker, American Academy of Forensic Sciences, February 2010.

[48] Brian Roux and Michael Falgoust. Ethical issues raised by data acquisition methods in digital forensics research. *Journal of Information Ethics*, 21(1):40–60, April 2012.

[49] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

[50] Stephen Yeazell. *Civil Procedure*. Aspen Publishers, seventh edition, 2008.

Appendices

Appendix A

Harris v BP - Order Compelling Discovery 5.1.1

BELINDA HARRIS, LYNN VINCENT, STALANAD, INC., MELBA TRAHAN & DAVID VINCENT	:	**DOCKET NO. 10-17687**
VERSUS	:	**38th JUDICIAL DISTRICT COURT**
BP AMERICA PRODUCTION COMPANY, CONOCO, INC., VASTAR RESOURCES, INC. TENNESSEE GAS PIPELINE CO. & RAE DONALDSON	:	**PARISH OF CAMERON**
	:	**STATE OF LOUISIANA**
FILED: _____	:	_____ **DEPUTY CLERK OF COURT**

ORDER

On October 7, 2009 a hearing was held on Plaintiffs' Motion to Compel Discovery against BP America Production Co. and its affiliate and predecessor companies (hereinafter collectively "BP").

After reviewing the briefs, motions, memoranda, and the other pertinent pleadings and documents of record in these proceedings;

IT IS HEREBY ORDERED, ADJUDGED, AND DECREED that all documents and other materials referenced in this Order be produced by BP within 10 days of the signing of this order;

IT IS FURTHER ORDERED that plaintiffs are entitled to have an attorney and an IT person (hereinafter "IT") present whenever any documents search is conducted in any database, file room, or other location where BP's documents are kept. The identification of any document discovered in these searches shall not constitute a waiver of any privilege. Privileges that may be asserted in good faith may be asserted at the time a document is identified and marked for copying by the plaintiffs;

IT IS FURTHER ORDERED that plaintiffs are entitled to conduct word and term searches that may have been conducted in the past regarding plaintiffs' discovery. The plaintiffs' IT is permitted to run searches on databases for data that may have been deleted over time;

IT IS FURTHER ORDERED that BP is to produce within the time as stated above any and all documents identified through discovery by Plaintiffs and listed on Page 2 of Plaintiffs' Motion to Compel Discovery (see Exhibit A attached to this order, which includes a list of documents);

IT IS FURTHER ORDERED that BP produce all documents relating to the reserve set for the Grand Chenier Gas Plant and Grand Chenier Separation Station as identified by Michael Haigood in his deposition on pages 102-105;

IT IS FURTHER ORDERED that BP produce all documents relating to the Grand Chenier Gas Plant and the Grand Chenier Separation Station that have not been produced thus far in discovery;

IT IS FURTHER ORDERED that BP produce all informative material, manuals, or research relative to NORM, saltwater disposal wells, lead, PCB, groundwater, produced water, pits, cooling towers, gas plants, or separation stations, including the manuals or research of Amoco and Arco;

IT IS FURTHERED ORDERED that BP produce the Arco assessment document similar to the North American Site Assessment identified by Michael Haigood in his deposition;

IT IS FURTHER ORDERED that BP produce all documents that pertain to past practices of defendants regarding the use of cement or synthetic pit liners, including, but not limited to, documents that identify the sites at which such liners were used, when the liners were installed, and any photographs of pits that contain or contained such liners;

IT IS FURTHER ORDERED that BP produce all soil and groundwater data that was collected at or near any gas plants or pits identified in the North American Site Assessment ("NASA") including, but not limited to: (a) the information identified on page 5 of June 1992 NASA - Current Status and Program Summary of the 18 gas plants assessed at facilities with operations that include NORM, Radium 226 and Radium; (b) all information on page 7 of NASA regarding programs to eliminate pits and replace them with tanks, (c) all accounting policies as identified on page 9 of NASA, (d) all the data and information pertaining to the three gas plants identified on page 14 of NASA, (e) all data, assessments and surveys regarding gas plants identified on page 17, (f) all data, assessments and surveys regarding as plants identified on page 18 of NASA, (g) all data and assessments/surveys regarding as plants identified on page 19 of NASA, (h) all data and assessments/surveys regarding gas plants identified on page 22 of NASA; and, (i) all risk evaluation forms of gas plants as identified in Appendix 6;

IT IS FURTHER ORDERED that BP produce all documents regarding current status and program summaries, including the final version of the North American Site Assessment;

IT IS FURTHER ORDERED that BP produce all research documents regarding the use of chromium as a corrosion inhibitor as identified in BP's letter dated April 3, 2009;

IT IS FURTHER ORDERED that BP produce all documents pertaining to environmental assessments, surveys, remediations and remediation costs of the Headlee gas plant including, but not limited to, the remediation of the groundwater regarding chlorides that migrated from a Chevron

plant next door;

IT IS FURTHER ORDERED that BP produce all documents pertaining to environmental assessments, and/or remediations of any gas plant or separation station, including, but not limited to, the following facilities: Zama Gas Plant, and Slaughter Gas Plant;

IT IS FURTHER ORDERED that plaintiffs are entitled to the: (1) EMS document and/or reports as identified in the EMS manual of the Grand Chenier Plant; (2) All process hazardous and analysis documents of the Grand Chenier Gas Plant and Separation Station; (3) All emails or any information regarding any environmental issues in any office of any employee who dealt in any way with the environmental conditions of the Grand Chenier Gas Plant and/or Separation Station.

IT IS FURTHER ORDERED that plaintiff, with the assistance of its own IT representative and an IT representative of BP, is permitted to access and search any available searchable databases that may contain relevant discovery material. BP is ordered to provide information as to how its documents and searchable databases are organized and configured. In addition, plaintiffs are entitled to conduct electronic searches of all available databases using the following search terms, or any other additional terms that may be reasonably derived from the use of the following search terms:

1. Gas Plant Separation Station
2. Produced Water Pit
3. Evaporation Pit
4. Reserve Pit
5. Burn Pit
6. Salt Water Disposal Well Breaches
7. Remediation of a Drinking Water Aquifer
8. Chromium Used as a Corrosion Inhibitor
9. Cooling Tower Problems
10. Waste Retention Pond
11. Salt Water Disposal
12. Information Regarding the Use of Salt Water Disposal Wells (when first used?)
13. Groundwater Remediation
14. NORM/Radium 226 and Radium 228
15. Produced Water Constituents

16. Pit Liners

17. Impact of Produced Water on Fish and Other Animals, and/or Plants;

IT IS FURTHER ORDERED that BP produce all editions and/or issues and/or versions of the Environmental and Social Review distributed by BP issued since its inception to the present (a copy of one such review is attached as Exhibit B for reference purposes);

IT IS FURTHER ORDERED that BP produce all environmental reports identified in the Environmental and Social Review document attached as Exhibit B, and all environmental reports identified in any other edition, issue or version of the Environmental and Social Review ; and

IT IS FURTHER ORDERED that BP produce all editions, issues or versions of all other BP in-house documents similar to the Environmental and Social Review.

IT IS FURTHER ORDERED that the court hereby defers its ruling on plaintiffs' request for penalties, attorneys fees and sanctions pending compliance by BP with the terms of this order.

Signed this 15th day of October, 2009.

Penelope Richard

JUDGE, 38TH JUDICIAL DISTRICT COURT

Appendix B

VPSB v. Louisiana Land, Letter from Plaintiff's Counsel 6.1.1

TALBOT, CARMOUCHE & MARCELLO

17405 PERKINS ROAD
BATON ROUGE, LOUISIANA 70810
TELEPHONE: (225) 400-9991
FAX: (225) 448-2568

DONALD T. CARMOUCHE*
VICTOR L. MARCELLO
JOHN H. CARMOUCHE

WILLIAM R. COENEN, III
JOHN S. DUPONT, III
BRIAN T. CARMOUCHE

*Member of the Texas Bar

AUBERT D. TALBOT
(1925-2005)

October 15, 2010

The Honorable Jerome M. Winsberg
5031 St. Charles Ave.
New Orleans, LA 70115

Re: State of Louisiana and the Vermilion Parish School Board
v. The Louisiana Land and Exploration Company, et al
Docket No. 82162; Div. D (East White Lake Field)
15th JDC; Parish of Vermilion; State of Louisiana

Dear Judge Winsberg:

On October 12, 2010, the Court heard argument concerning plaintiffs' motion to compel access to Unocal's document database. As a temporary measure, the Court ordered that the list of search terms run by Unocal be provided to plaintiffs for further review. Plaintiffs have begun a thorough review of this list, but it is already apparent that the searches were inadequately designed. As a result, it appears that Unocal's search is likely to have missed some crucial documents that relate directly to this particular piece of property and this lawsuit.

At the outset, it is critical to understand that the adequacy of the database searches run by Unocal is *the* key determinant of whether Unocal's production is complete in this case. In a typical discovery production, some file clerk or records custodian who is familiar with the company's files works with counsel to gather all of the files that might contain responsive documents. Those files are then reviewed by counsel and any responsive documents are produced. The key to the reliability of the typical model is that the knowledgeable file clerk or records custodian can testify that, "Our review included all of our files that might have contained responsive documents." In this case, unlike the typical situation, there is no one at Unocal who can make this key statement.

There is no single shelf, file room, or warehouse which Unocal can point to and say, "Here are all of our files on the East White Lake Field." The files (or indexes of them) have been digitized and scattered among Unocal's various database servers, with the hard copies held by third-party off-site storage companies. Because these files can only be "gathered" for review by running database searches (as opposed to walking over to a particular shelf in a file room), the quality of these database searches is the key to whether or not all of the pertinent files are reviewed for responsive documents. Regardless of whether Unocal and their counsel did a first-rate review and produced every single responsive document from the files they reviewed, Unocal's production would be incomplete if the database searches were not successful in

identifying all of the files that should be reviewed.

Because the success of database searches is so critical, it's also important to understand a little about how they work. Unlike Google searches which use complex algorithms to find results which are "close" matches to the search terms, Unocal's databases, as far as we can tell, run Boolean-type searches. Boolean searches produce very precise results, but they are only as good as the search terms used. For instance, a search for "pits" would return every file whose indexed description includes the word "pits," but would not return files indexed with the word "pit" (unless the index also included "pits"). Likewise, a search for "unocal AND pit" would only return results where the words "unocal" and "pit" appear in the same index. The rifle-shot precision of Boolean searches can also lead to omissions that are not readily apparent. One might expect a search for "(Pit or Pits) AND "East White Lake"" to return all files dealing with pits in the East White Lake Field; however, if a file was indexed as "Union Oil Pit Remediation File for E. White Lake Field," it would not be included in the results because the exact phrase "East White Lake" was not used when the file index was created. Because plurals, abbreviations, misspellings and other variations in the indexes can have such an important effect on the results of a search, the best method is to cast a wide net first, then narrow it down to exclude irrelevant files.

Upon reviewing Unocal's list of searches, it is obvious that only a very small and very selective net was cast. At the very least, we expected that Unocal would have reviewed all of its files indexed with "VPSB" or "East White Lake" for responsive documents. To our surprise, we found that except with respect to Carrolton Resources files, Unocal only ran searches for "East White Lake" and "VPSB" in combination with other terms such as "(sample or sampling)," "(study or studies)," "assessment," "survey" and "test." This is completely unnacceptable considering that a file indexed as "Phase I & II Env. Assessments of SW Pit - VPSB Lease - East White Lake Field" would not have been included (because the search did not include the plural of "assessment"). Even more damning of Unocal's inadequate searches is the fact that they would have missed a file indexed as "Unocal Liability for Groundwater Contamination from Leaking Pit – VPSB Lease - East White Lake Field." How much more on point could a file possibly be? Yet if that file existed, Unocal would not have produced it because they strategically chose not to cast too wide a net with it's database searches.

These are just a few of the initial problems that plaintiffs see with the searches performed by Unocal. However, we feel that these alone are enough to show that, it is highly likely that important documents could be missing from Unocal's production in this case. For these reasons, plaintiffs will wish to reurge their motion to compel and present additional argument and testimony from one of plaintiffs' IT experts on October 25, 2010.

With kindest regards, I am

Very truly yours,

William R. Coenen, III

WRCIII/jke
Attachments
cc: Grady J. Abraham
Alan J. Berteau/L. Victor Gregoire
Michael R. Phillips
K. Wade Trahan
Carol M. Wood/Robert E. Meadows
Calvin E. Woodruff, Jr.

Appendix C

VPSB v. Louisiana Land, Plaintiffs' Brief in Support 6.1.3

15TH JUDICIAL COURT FOR THE PARISH OF VERMILION

STATE OF LOUISIANA

DOCKET NO. 82162 **DIVISION "D"**

STATE OF LOUISIANA and THE VERMILLION PARISH SCHOOL BOARD

VERSUS

THE LOUISIANA LAND AND EXPLORATION COMPANY, PEAK OPERATING CO., UNION OIL COMPANY OF CALIFORNIA, UNION EXPLORATION PARTNERS, LTD., CARROLLTON RESOURCES, L.L.C., AND PHOENIX OIL & GAS CORPORATION

FILED: _____ **_____**
DEPUTY CLERK OF COURT

PLAINTIFFS' SUPPLEMENTAL BRIEF IN SUPPORT OF MOTION TO COMPEL ACCESS TO DATABASE AND RULE 9.8 STATEMENT

NOW INTO COURT, through undersigned counsel, come plaintiffs, who file this supplemental brief in support of their motion to compel access to the Chevron/Unocal/Carrollton database.

On October 12, 2010, the Court ordered that defendants produce a list of the terms that they used to conduct searched of their informal database. A review of these search terms has confirmed that plaintiffs' search for documents in this case was woefully inadequate and likely missed crucial documents.

At the outset, it is critical to understand that the adequacy of the database searches run by Unocal is *the* key determinant of whether Unocal's document production is complete in this case. In a typical discovery production, a file clerk or records custodian who is familiar with the company's files works with counsel to gather all of the files that might contain responsive documents. Those files are then reviewed by counsel and any responsive documents produced to the requesting party. The key to the reliability of the typical model is that the knowledgeable file clerk or records custodian can testify to the effect that, "I personally performed this search and my review included all of our files that might have contained responsive documents." In this case, unlike the typical situation, there is no one at Unocal who can make this key statement.

There is no single shelf, file room, or warehouse that Unocal can point to and say, "Here

are all of our files on the East White Lake Field." The files (or indexes of them) have been digitized and scattered among Unocal's various database servers, with the hard copies held by third-party off-site storage companies. Because these files can only be "gathered" for review by running database searches (as opposed to walking over to a particular shelf in a file room), the quality of these database searches is the key to whether or not all of the pertinent files are reviewed for responsive documents. Regardless of whether Unocal and their counsel did a first-rate review and produced every single responsive document from the files they reviewed, Unocal's document production would be incomplete if the database searches were not successful in identifying all of the files that should be reviewed.

Because the success of database searches is so critical, it is also important to understand a little about how they work. Unlike Google searches which use complex algorithms to intuit close keyword matches and which identify documents that are close to, rather than exactly, the search terms used, Unocal's databases, as far as we can tell, run strict Boolean searches. Strict Boolean searches produce very precise results, but their effectiveness is limited by the quality of the search query used. For example, a search for "pits" would return every file whose indexed description includes the word "pits," but would not return files indexed with the word "pit" (unless "pits" was also present); likewise a search for "salt water" would pass over "saltwater" and "salt-water" as valid matches. The rifle-shot precision of strict Boolean searches can also lead to omissions that are not readily apparent. One might expect a search for "(Pit or Pits) AND "East White Lake"" to return all files dealing with pits in the East White Lake Field; however, if a file was indexed as "Union Oil Pit Remediation File for E. White Lake Field," it would not be included in the results because the exact phrase "East White Lake" was not used when the file index was created. Because plurals, abbreviations, misspellings and other variations in the indexes can have such an important effect on the results of a search, the best method is to cast a wide net first, then narrow it down to exclude irrelevant files. *See* Affidavit of Brian Roux.

Upon reviewing Unocal's list of searches, it is obvious that only a very small and very selective net was cast. At the very least, we expected that Unocal would have reviewed all of its files indexed with "VPSB" or "East White Lake" for responsive documents. To our surprise, we

found that except with respect to Carrollton Resources' files, Unocal only ran searches for "East White Lake" and "VPSB" in combination with other terms such as "(sample or sampling)," "(study or studies)," "assessment," "survey" and "test." This is completely unacceptable considering that a file indexed as "Phase I & II Env. Assessments of SW Pit - VPSB Lease - East White Lake Field" would not have been included (because the search did not include the plural of "assessment"). Even more damning of Unocal's inadequate searches is the fact that they would have missed a file indexed as "Unocal Liability for Groundwater Contamination from Leaking Pit – VPSB Lease - East White Lake Field." Yet if that file existed, Unocal would not have produced it because they strategically chose not to cast too wide a net with their database searches. Further, plaintiffs aver that, especially under the time constrains of the trial schedule for this matter, plaintiffs cannot prepare a list of search terms which would adequately address these concerns without access to perform real-time searches on the Chevron/Unocal/Carrollton database. The design of successful search parameters is largely dependent upon rial and error due to the facts that (i) search results are dependent upon the way the files are indexed, and (ii) database query "grammar" varies substantially among the particular database software packages. Without the ability to perform real-time searches, get immediate feedback in the form of search results, and adjust the search terminology in order to obtain the most complete results, the process of designing and performing adequate searches, processing the results, reviewing the underlying hard copy files and producing the responsible documents cannot be completed in time for plaintiffs to conduct a meaningful review before trail. *See* Affidavit of Brian Roux.

These are just a few of the initial problems that plaintiffs see with the searches performed by Unocal. However, plaintiffs feel that these alone are enough to show that it is highly likely that important documents could be missing from Unocal's production in this case. *See* Affidavit of Brian Roux. For these reasons, plaintiffs wish to re-urge their motion to compel and present additional argument and testimony from one of plaintiffs' IT experts on October 25, 2010.

RULE 9.8 STATEMENT

Pursuant to Rule 9.8 of the Uniform Rules for District Courts, plaintiffs may offer live testimony on October 25, 2010 in support of their Motion to Compel Inspection of the Chevron/Unocal/Carrollton Database. This case is set for trial beginning November 15, 2010.

Respectfully submitted;

TALBOT, CARMOUCHE & MARCELLO
17405 Perkins Road
Baton Rouge, LA 70810
(225) 400-9991 Telephone
(225) 448-2568 Fax

_____
Donald T. Carmouche (La. Bar #2226)
Victor L. Marcello (La. Bar #9252)
John H. Carmouche (La. Bar #22294)
William R. Coenen, III (La. Bar #27410)
John S. DuPont, III (La. Bar #26271)
Brian T. Carmouche (La. Bar #30430)

Grady J. Abraham (La. Bar #24739)
ATTORNEY AT LAW
120 E. Third St.
P.O. Drawer 2309
Lafayette, LA 70502-2309
(337) 234-4523 Telephone
(337) 234-4547 Fax

Calvin E. Woodruff, Jr. (La. Bar #13666)
ATTORNEY AT LAW
111 Concord St., Ste. A (70510)
P.O. Box 520
Abbeville LA 70511-0520
(337) 898-5777 Telephone
(337) 898-5781 Fax

Attorneys for Plaintiffs

15TH JUDICIAL COURT FOR THE PARISH OF VERMILION

STATE OF LOUISIANA

DOCKET NO. 82162 **DIVISION "D"**

STATE OF LOUISIANA and THE VERMILLION PARISH SCHOOL BOARD

VERSUS

THE LOUISIANA LAND AND EXPLORATION COMPANY, PEAK OPERATING CO., UNION OIL COMPANY OF CALIFORNIA, UNION EXPLORATION PARTNERS, LTD., CARROLLTON RESOURCES, L.L.C., AND PHOENIX OIL & GAS CORPORATION

CERTIFICATE OF SERVICE

I HEREBY CERTIFY that a copy of the above and forgoing pleading has been forwarded to the following counsel of record by placing same in the United States mail, postage prepaid, and properly addressed:

William G. Jarman
Jeffrey N. Boudreaux
L. Victor Gregoire
Alan J. Berteau
KEAN, MILLER,
HAWTHORNE D'ARMOND,
McCOWAN & JARMAN, LLP
One American Place 22nd Floor
P.O. Box 3513
Baton Rouge, LA 70821
Attorneys for Union Oil Company of California, Union Exploration Partners, Carrollton Resources LLC, Chevron U.S.A., Inc., Chevron Midcontinent, LP

Michael R. Phillips
KEAN, MILLER,
HAWTHORNE D'ARMOND,
McCOWAN & JARMAN, LLP
First Bank and Trust Tower
909 Poydras St., Suite 1400
New Orleans, LA 70112
Attorney for Union Oil Company of California, Union Exploration Partners, Carrollton Resources LLC, Chevron U.S.A., Inc., Chevron Midcontinent, LP

K. Wade Trahan
OTTINGER HEBERT, L.L.C.
1313 West Pinhook Road
P.O. Drawer 52606
Lafayette, LA 70505-2606
Attorney for Union Oil Company of California, Union Exploration Partners, Carrollton Resources LLC, Chevron U.S.A., Inc., and Chevron Midcontinent, LP

Carol M. Wood
Robert E. Meadows
Shelby E. Wilson
KING & SPALDING, LLP
1100 Louisiana, Ste. 4000
Houston, TX 77002
Pro Hac Vice Counsel for Union Oil Company of California, Union Exploration Partners, Carrollton Resources LLC, Chevron U.S.A., Inc., and Chevron Midcontinent, LP

Baton Rouge, Louisiana this _____ day of October, 2010.

_____
William R. Coenen, III

Appendix D

VPSB v. Louisiana Land, Defendants' Opposition 6.1.4.2

15TH JUDICIAL DISTRICT COURT

PARISH OF VERMILION

STATE OF LOUISIANA

STATE OF LOUISIANA AND THE VERMILION PARISH SCHOOL BOARD	DOCKET NO. 82162
VERSUS	DIVISION: D
LOUISIANA LAND AND EXPLORATION COMPANY, ET AL.	JUDGE: WINSBERG (AD HOC)

DEFENDANTS' MEMORANDUM IN OPPOSITION TO PLAINTIFFS' SUPPLEMENTAL BRIEF IN SUPPORT OF MOTION TO COMPEL ACCESS TO DATABASE

As predictably as snow in Alaska, Plaintiffs profess themselves disappointed in the search term list provided to them by UNOCAL, as the Court ordered. Just as counsel for UNOCAL noted at the October 12, 2010 hearing, it is inconceivable that Plaintiffs will ever admit to satisfaction with this, or any, search term list proffered by a defendant in an oilfield legacy case. This is particularly so where, as here, their claimed dissatisfaction furthers their frantic efforts to have the November 15, 2010 trial date continued. Plaintiffs' consultant, Brian Roux, has never met a defense search term list that he did not characterize as "woefully inadequate" – because that it is his job. His job is to generate discovery disputes so that Plaintiffs, instead of being relegated to the distasteful task of proving their case on the merits, can run amok in their opponents' computers and databases.

Plaintiffs have quietly discarded their complaint that UNOCAL failed to search for an "Alpha Environmental" report. Through their "supplementation," Plaintiffs have essentially filed a brand new motion to compel, without the requisite conference required by Rule 10.1 of the Uniform Rules for Louisiana Civil District Court. Plaintiffs obviously wish to have a placeholder motion they can periodically "supplement" as the mood strikes them. The Louisiana Code of Civil Procedure does not operate so loosely. On October 12, 2010, the Court heard, and ruled upon, Plaintiffs' motion to compel access to Defendants' databases. The Court ordered UNOCAL to produce its list of search terms and it did so, in open court. One business day before the next-scheduled hearing on

October 25, Plaintiffs spring an entirely new motion, drawing upon the purported expertise of Mr. Roux, and meticulously deconstructing UNOCAL's search term list. This motion should be denied, and Plaintiffs should be required to file a proper motion to enable UNOCAL to defend itself from the sweeping implications of the new motion - including, but not limited to, deposing Mr. Roux about the factual basis for Plaintiffs' allegations.

Plaintiffs propose a "principle" here that has no place in the law. If a party deems itself dissatisfied with the document production of an opposing party, it need present no objective evidence to support that dissatisfaction – such as, for example, actual **proof** of the existence of a document which has not yet been produced. Instead, the articulation of this dissatisfaction suffices to entitle the complaining to disregard the **actual production** to date, usurp the position of the opposing party's counsel, and dictate anew the scope of discovery. This is not now, and has never been, the law in this State. Plaintiffs cite no support for their position. They would have the Court accept the "principle" because they say it is so.

Plaintiffs argued in their original motion to compel that an "Alpha Environmental" report was missing, and should be produced. This was the stated justification for their desire to invade UNOCAL's database. Plaintiffs have now quietly discarded these grounds, and instead posit the existence of an **imaginary document** -- "Union Oil Pit Remediation File for E. White Lake Field" – which UNOCAL's searches would presumably not have identified. Plaintiffs are playing opportunistic word games with this Court and with the law. Furthermore, documents covering pit remediation are readily available through the Louisiana Department of Natural Resources, and UNOCAL produced, in July 2010, over 200 pages of pit files for the property at issue. Once again, Plaintiffs fail to identify any document that is genuinely missing.

Aside from Plaintiffs' disregard for procedure and fair notice, Plaintiffs and Mr. Roux would have the Court ignore the painstaking searches conducted by UNOCAL in this case in 2004, 2005 and 2006, before the onset of "complex algorithms" or "Boolean searches." UNOCAL's counsel looked through UNOCAL's archived East White Lake

documents in 2004, 2005 and 2006, before these documents were merged into Chevron's TIMES database. UNOCAL's counsel looked through UNOCAL's archived East White Lake documents in 2004, 2005 and 2006, before these documents were merged into Chevron's TIMES database. The results of those searches, over 15,000 pages of documents, were produced.

Plaintiffs and Mr. Roux would also have the Court to ignore the 22,000 pages of documents produced by Peak Energy, the oil and gas operator on the property at issue for the last 15 years. The majority of those documents are of UNOCAL originals, bequeathed to Peak, through its predecessor, Resource Acquisitions, by UNOCAL in 1995.

Database searches have never been the "key determinant" of the completeness of document production in this case. In the earlier stages of the case, UNOCAL's lawyers were, in fact, required to look through all of the boxes which might contain documents relevant to this lawsuit. They went to warehouses and storage facilities to find, and review, all of the available documents relating to the East White Lake Field. There were no electronic indices, no "search terms," no "complex algorithms," no "Boolean searches" used. UNOCAL's had ceased operations at East White Lake Field almost a decade earlier. The documents in question were archived on shelves, in boxes with labels. The boxes with labels that said "East White Lake" were searched. This approach may not be sophisticated, but at least it does not readily lend itself to the word games employed here by Plaintiffs.

Finally, Plaintiffs ignore the fact that a box identified in TIMES through use of the phrase "East White Lake" – the key phrase Plaintiffs say should have been used by UNOCAL in its searches -- would also be identified through one or more of the search terms used by UNOCAL in this case. TIMES is a folder-level index. If the TIMES description of one folder in a given box contains the word "Vermilion," and another contains the word "UNOCAL," the box would be identified. Mr. Roux does not note this probability. Of course, any knowledge Mr. Roux lacks about the TIMES database, and how it operates, lies at Plaintiffs' doorstep. They had an opportunity to question UNOCAL's 1442 witness about how the TIMES database operates, and they failed to do

so. They should not now be allowed to deploy this ignorance as an offensive weapon, by failing to note the obvious flaws in their own argument.

As UNOCAL noted in its original brief, Plaintiffs' suspicion – raised for the first time on the eve of trial -- that they have not received all of the documents in UNOCAL's possession to which they are entitled is not only groundless, but cannot serve as the basis for the burdensome sanction Plaintiffs seek. *Ford Motor Co. v. Edgewood Properties, Inc.*, 257 F.R.D. 418 (D.N.J. 2009). There is simply no reason to re-start discovery in this case. Plaintiffs are obviously not ready for trial, and are looking for any pretext they can seize to avoid facing the consequences of having a meritless case. Plaintiffs have shown no wrongdoing on UNOCAL's part which would support any sanction, much less the profoundly disruptive sanction Plaintiffs propose. Plaintiffs' new motion to compel – masked as a "supplemental brief" to their previous motion – should be denied, both on procedural and substantive grounds.

Respectfully submitted:

G. William Jarman, #7238
L. Victor Gregoire, #22400
Alan J. Berteau, #17915
KEAN, MILLER, HAWTHORNE, D'ARMOND, MCCOWAN & JARMAN, L.L.P.
Post Office Box 3513
Baton Rouge, LA 70821
Phone: 225.387.0999
Fax: 225.388.9133

K. Wade Trahan (La. Bar No. 20474)
OTTINGER HEBERT, L.L.C.
1313 West Pinhook Road (70503)
Post Office Drawer 52606
Lafayette, Louisiana 70505-2606
Telephone: (337) 232-2606
Fax: (337) 232-9867

Robert E. Meadows, *Pro Hac Vice*
Carol M. Wood, *Pro Hac Vice*
Shelby E. Wilson, *Pro Hac Vice*
KING & SPALDING
1100 Louisiana Street - Suite 4000
Houston, TX 77002-5213
Phone: (713) 751-3200
Fax: (713) 751-3290

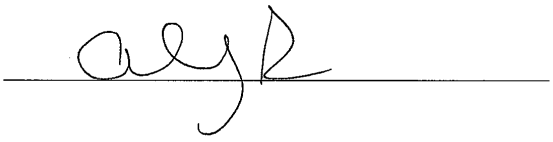
Attorneys for Chevron Midcontinent, LP, Union Oil Company of California, Union Exploration Partners and Carrollton Resources, LLC

CERTIFICATE OF SERVICE

I hereby certify that a copy of the foregoing document has been sent via U.S. Mail, postage prepaid and properly addressed, to all known counsel of record.

John H. Carmouche Donald T. Carmouche Victor Marcello William R. Coenen III TALBOT, CARMOUCHE & MARCELLO 17405 Perkins Road Baton Rouge, LA 70810	Mr. Calvin E. Woodruff, Jr. Legal Counsel/Risk Manager Vermilion Parish School Board P.O. Box 520 Abbeville, LA 70511-0520
Grady J. Abraham Attorney at Law 120 E. Third St. P.O. Drawer 2309 Lafayette, LA 70502-2309	

Baton Rouge, Louisiana this 29TH day of October, 2010.



Appendix E

Transcripts November 3, 2010

forensic IT searches?

A. About five years.

Q. And that's exclusively IT investigations, right?

A. Yes.

MR. COENEN: At this point, Your Honor, we'll offer Mr. Roux as an expert in the field of forensic IT investigations. We tender the witness.

MR. BERTEAU: We're talking about qualifications at this point? I just have probably one question.

EXAMINATION BY MR. BERTEAU:

Q. Mr. Roux, have you ever had occasion to work with the Times Electronic Index at issue in this case?

A. No.

MR. BERTEAU: Your Honor, I have no basis upon which to dispute his qualifications. I merely would like the right to reserve -- I mean, reserve the right to explore the particular application of that expertise in this case.

THE COURT: He is qualified as an expert in the field offered.

MR. COENEN: Thank you, Your Honor.

EXAMINATION BY MR. COENEN:

Q. You've been searched -- you've been provided a list of search terms for the searches that were run of the Chevron Times database in this case, have you not, Mr. Roux?

A. Yes, I have.

MR. COENEN: I will offer -- we'll mark as Plaintiff Exhibit 1.

EXAMINATION BY MR. COENEN:

Q. A copy of the search terms that were provided entitled privileged and confidential attorney work product and confidential attorney client information that was handed to us a

hearing or two ago concerning the search terms, and just ask if you can identify that, please, as the list that you actually reviewed.

MR. COENEN: Your Honor, may I use this podium?

THE COURT: Yes.

A. (The witness examines the document.) Yes. This is the list I reviewed.

EXAMINATION BY MR. COENEN:

Q. Sir, that's the actual list that you reviewed?

A. Yes.

Q. First of all, from looking at that list of search terms, can you tell the Court a little bit about the type of database that appears that Chevron or Carrollton or UNOCAL is actually running?

A. The method of the search queries is a Boolean search, which is very common in these sort of document systems. Usually these are referring to physical files, physical boxes of files from legacy kind of document productions, as opposed to more modern systems which would store electronic versions of the files. Typically they -- or at least from my experience of what this looks to be searching, is probably something along the lines of what a commercial product would be, like concordance where the data is stored in fields and they're searching with the Boolean searches for specific key terms that the physical boxes have been indexed with.

Q. Tell the Court -- if I could stop you right there. Tell the Court what a Boolean system, B-o-l -- B-o-o-l-e-a-n system?

A. Boolean comes from a formal logic system in which statements are evaluated as true and false and in combinations of ands and ors and negations. It results in a yes or no to the search query. If it returns yes, then the document is returned;

if it returns no, then the document is not returned.

Q. You indicated that this type of database being run by Chevron appears to be a Boolean-type system, right?

A. Correct.

Q. Now, you've been involved in these types of cases before, and when you mentioned the *Harris* case earlier, and you're aware of the types of documents, are you not, that the plaintiffs were looking for from Chevron and Carrollton and UNOCAL in this case, correct?

A. Yes.

Q. You understand that they're basically seeking to obtain, at a minimum, all of the records concerning the East White Lake Field, right?

A. Correct.

Q. Which is in Vermilion Parish which is related to this case, correct?

A. Correct.

Q. Explain to the Court, if you would, why that list of document search terms that I've identified as Plaintiff's Exhibit 1 would not, in fact, return all of the documents having to do with the East White Lake Field.

A. To begin with, if you look at -- many of the search terms, such as "pits," which is a common example that has come up before, is searching in the singular rather than the singular and the plural. It doesn't seem to use features of these sorts of databases that have root expansion to where a shortened form of the word can expand into multiple derivations of it. If the system doesn't support that it can be mimicked using more complex Boolean expressions, different "ors" for example. If it didn't support searching both plural and singular from the root expansion you could do it by saying "and" and using a parenthesis type expression pit or pits, et cetera. So many of

these not only do not search the plural and singular, they also have separate searches for, say, "saltwater" and then "salt water" as a literal with a space, and there seems to be some duplication. There's a lot of potential for duplication of the documents that are produced because they are searched individually; whereas, if you were using a properly formed search query you would return the documents only one time as opposed to multiple times.

Also, there is abbreviations that I was asked to look at, which I don't recall the specifics, but that were not present in the search queries that were provided, and without that the document did not or the document index did not refer to the specifically, say, Vermilion as opposed to some abbreviation thereof, then that would not be returned because these indexes are not full text indexes.

Q. Let me stop you right there. I gather from what you're saying is either the searches were too simplistic or not simplistic enough in some cases, correct?

A. Yes.

Q. You said as far as abbreviations go. Vermilion Parish School Board, as you know, is sometimes abbreviated VPSB?

A. Yes.

Q. In fact, you found that acronym quite often in the case, did you not?

A. Yes.

Q. Did the searches that they run include a search for just VPSB by itself?

A. No, they didn't. In fact, none of their searches were singular words, except for "pits," that I see offhand. And the main issue with that is the nature of the database being that these are old documents. As I understand it because they're not a full text index it's more likely that an abbreviation is going

to be used in a field that has a limited space to store information. And so without searching for those abbreviations in general you run the risk of overlooking documents, especially in cases where almost everything is conjoined with other terms.

Q. Let me ask you this. Did they search for the words "East White Lake" independently by themselves -- the phrase "East White Lake"?

A. No.

Q. That was not searched for?

A. No.

Q. Is that something that you think as a Forensic IT specialist that a company like Chevron or UNOCAL or Carrollton, in an environmental case dealing with the East White Lake oil and gas field, should have looked for?

A. Yes. I think if the search is going to be conducted in a more methodical fashion that you would start with a much more general case to pull all the documents that have the specific words then you would cull it down from there, based on Louisiana, for example, once you had pulled everything in the off chance that Louisiana was not present just to see the differences in the documents produced.

Q. And you also indicated I think, and I want to make sure this is clear for the Court, that they limited the effectiveness of their searches by improper use of plural search terms, correct?

A. Correct. Especially in a case where it's not a full text of the document that you're searching but just rather just an index that has been entered by someone to describe the document. The singular or the plural version is likely to be used in and of itself, just by itself rather than occur in both formats throughout the document text that you would find in a full text index.

Q. Is there anything else that you see that was improper with respect to their searches, sir?

A. In general, in terms of what's produced, I would say that the duplication of documents are likely to be resulting from this type of searching is probably going to create a more burden for reviewing the documents, but there's some issue with field information. I'm not sure from the list that they're showing whether this is searching all of the fields, whether this is searching a specific text field, whether this searches the individual, like, if Union Oil appeared in one field and Louisiana appeared in a different field if it would pick that up or whether it would only pick it up if it occurred in the same field.

Q. Is this the sort of thing that you could pick up quickly if you were actually there looking at their database?

A. Yes. Generally, these sort of systems are fairly similar in the way that they present themselves from a user interface perspective and from a query language perspective. So usually within a few queries you can determine how exactly this system is behaving.

Q. In fact, you did that in the *Harris* case; you were able to get visual inspection of the database and get real-time feedback for how your searches were running, correct?

A. Correct. We did that both at BP's discontinued operations division and Conoco's world headquarters.

Q. That was with Conoco and with BP's database systems?

A. Correct. And they --

Q. And one of the firms involved in that case was Mr. Alan Berteau's firm, correct?

A. Yes, actually.

Q. And he was involved in some of the searches of those documents, as well?

A. Yes. Alan and I along with Todd were in the room there when we were running the searches.

Q. And the method that you're proposing to use in this case to search the Chevron database would be exactly the same method you would propose we should use -- I'm sorry, the method proposed to use in this case was the same method that was actually used to search the Conoco and BP databases?

A. That's correct.

Q. And you were able to do that even with lawyers in the room and it was a well supervised activity, was it not?

A. Yes. It seems any time there was an issue that occurred the two sides could work out the issue. I don't think we had to resort to calling outside at any point during it. I think there was a few incidents for them to consult with each other's people back home, but I don't think there was any sort of ruling or anything of that nature, to my knowledge.

Q. In other words, it was done with minimal fuss; you didn't have to call the judge in other words?

A. It was very cordial. Yes.

Q. In that particular case were you able to find more documents in the BP and Conoco databases that pertain to that case and actually had been produced by the other side?

A. Yes. In the BP case, I believe, we pulled up one document it was indicated didn't exist. In the Conoco case we actually ran into a completely different issue of user privileges within the database that we found that they had originally been searching using a user that had insufficient privileges to search the entire database, and once the super user had been provided the number of search results increased exponentially.

Q. If you were to do that same sort of thing in this case, would it help you, in your opinion, to discover and

confirm whether or not what Chevron is telling us they produced in this case is, in fact, all the documents?

A. Can you state that again?

Q. Fair enough. If you were to perform this sort of search, this in-house search with real-time feedback of their system, would that help you to determine whether or not all the documents, in fact, concerning the East White Lake Field had been produced?

A. Yes. Yes, it would.

MR. COENEN: Thank you, Mr. Roux.

EXAMINATION BY MR. BERTEAU:

Q. Mr. Roux, as I appreciate it, your source of your dissatisfaction with the searches that were run in this case would fall under two headings, duplicative and failed to run certain searches that you think should have been run; is that accurate?

A. That's accurate, yes.

Q. You would agree, would you not, that the running of duplicative searches would in essence have the affect of producing boxes over and over again, cause more work for us in reviewing those documents. but ultimately wouldn't miss documents, is that correct?

A. It depends. With the searches that are duplicative you would end up with duplicate record entries; whether you pulled those boxes multiple times would be a different issue. But it would not pass over documents from the duplicative angle.

Q. In other words, we would be alerted to them, and if we pulled the boxes we'd review and we'd find them?

A. Uh-huh (affirmative response.)

Q. So in terms of identifying documents, duplicativeness, if you will, is not a problem; is that accurate?

A. It's not a problem -- well, from the perspective of

what the end result is, but it is symptomatic of the person crafting the searches may not have had an appropriate level of expertise.

Q. But it's not a problem with the search term; it's a problem with the human response to the search term?

A. Say that again.

Q. It's a problem with the human response, the human implementation of the results.

A. Yes, the implementation would be the correct way to put it.

Q. How long have you worked with the Carmouche firm?

A. The first case I dealt personally with them was the *Harris* and *BP.* and that's the only case I've actually dealt with them.

Q. The *Harris* case, as I recall, was December of 2009?

A. That's correct. Yes.

Q. That case had been around for a number of years before you got involved, is that right?

A. I don't know.

Q. How long has your firm been involved in this case?

A. I don't know specifically. I don't deal with the other aspects and what cases they handle.

Q. How long have you been involved in this case?

A. Two weeks, three weeks, something of that nature. I don't recall the exact date.

Q. In other words, your first involvement in the case was when you were asked to review the list of search terms that Mr. Coenen referred to, right?

A. That is correct.

Q. Did you review the actual discovery propounded on UNOCAL by the plaintiffs in this case?

A. Could you be --

Q. Did you review the written discovery?

A. No.

MR. BERTEAU: Your Honor, I'm going to mark as Defense Exhibit 1 a copy of plaintiffs first set of discovery requests to defendants. I think it's attached to the motion, and just to confirm -- May I approach the witness, Your Honor? I'm sorry. Gallivanting around here.

THE COURT: Yes.

EXAMINATION BY MR. BERTEAU:

Q. Have you seen this document before, Mr. Roux?

A. No.

Q. Have you seen any other written discovery actually propounded in this case?

A. The only other document I reviewed was a transcript from a deposition of -- I don't remember the individual's name. He was in a systems role at UNOCAL or Chevron or something like that.

Q. A defense witness? Plaintiff witness?

A. It was a defense witness testifying -- or being deposed about the time system.

Q. Did you play any part -- I gather you did not play any part in the drafting of the discovery -- plaintiff's discovery in this case; is that right?

A. No, I did not.

Q. At no time during this case have you ever been asked to or drafted a request -- search terms to request UNOCAL to run; is that correct?

A. No, I have not.

Q. Did you have occasion or did you review any of the documents actually produced by UNOCAL in this case?

A. No.

Q. So if, for example, I were to say to you UNOCAL produced several hundred pages of documents relating to pit closures, you would not have a basis to disagree with that; would you?

A. No.

Q. You would not have a basis to disagree with the conclusion that whether we used the search term "pit" or "pits" we found the pit files in this case; you don't have any basis to disagree with that, do you?

A. No.

Q. Same with saltwater documents, documents relating to saltwater production or disposal on the East White Lake Field, did you look to see if any of those documents had actually been produced?

A. Again, I haven't reviewed any of the produced documents.

Q. Mr. Roux, would you agree that in order to determine the effectiveness of a given set of search terms you need to see both the discovery propounded and the actual documents produced in response to that discovery?

A. Say that again.

Q. Let's break it down. Wouldn't you agree that in order to design effective search terms you'd have to look at the discovery that your client is propounding on the other party?

A. Well, in this particular case the question that I was asked was to review the nature of the search queries themselves, their structure and how they were designed. To review the effectiveness of these versus a hypothetical better version of them. I would have to look at both the documents produced from these as well as documents produced from the proposed search queries to see what the difference between the two sets were.

Q. In order to gauge the effectiveness of the search

terms listed by UNOCAL, wouldn't you have to see the discovery propounded by the plaintiffs to UNOCAL?

A. It would also require something to compare it against in order to determine whether these or how effective these were versus an alternate version of the search queries.

Q. But you would agree that a review of the actual discovery would be important in that comparison?

A. Yes.

Q. And you didn't do that this case?

A. No.

Q. Again, in order to gauge the effectiveness of the actual search terms used, wouldn't you agree that you'd have to look at the actual documents produced by the parties who ran the search terms?

A. It depends on how you refer to the effectiveness. If you're talking about in a specific instance, it's possible reviewing the documents could on the individual search queries determine the effectiveness. However, because the search queries are searching the indexes and not the full text of the documents, the text of the documents is not very helpful to determining the effectiveness of this because it is the indexes that are returned that are relevant to the search queries. So it's more -- it would be more effective to look at the number of documents returned and the overlap from these versus an alternate set of search queries.

Q. If you have discovery requests that seek documents relating to pits and you have documents produced that related to pits, don't you have to look at the documents produced in response to that discovery request to determine whether or not the defendant or the party searching for those documents is at least implementing search terms that find relevant documents?

A. You would look at the doc -- could you repeat that

again?

Q. It's back to my original question. I want to make sure I have a clear answer to it, if I can. Doesn't the nature of the documents produced, i.e., you say we should have ran "pit," we ran "pits," whatever the case may be, in addition to the other search terms, if, in fact, we produced hundreds of pages of documents relating pits, haven't we generated the documents responsive to the discovery?

A. Not necessarily. Because the search terms are not running against the full text of the documents, this would be a different case if these were like Word documents or something in which the entire document was searched, but these are indexes that as I understand it are entered by someone at the time that the physical documents are put into storage. If you run, for example, "pits," since that seems to be the easiest example that we like to go back and forth with, you search for everything that has "pit" or "pits" or derivations thereof, you're pulling the indexes that somebody has entered about the physical document. That doesn't necessarily mean that all of the physical documents that involve "pit" or "pits" are going to be returned by that search query, if that makes sense.

Q. It does. You said before you're not familiar with the Times index are you?

A. No.

Q. It's possible, is it not, that if you run pits you could capture variations of the word "pits"?

A. It's possible, but that would seem unlikely given the way that these search terms are structured.

Q. Sitting here today, do you know of the existence of any document in this case that someone has said exists but has not been produced?

A. No.

THE COURT: How can you know that unless you run the way you want to run it as opposed to the way it's been run?

THE WITNESS: You wouldn't be able to determine if something hadn't been produced.

THE COURT: You believe if you ran yours you're liable to find some that haven't been produced?

THE WITNESS: The first -- Running alternate search queries and comparing the results against what has already been produced from an index doesn't have to produce the actual documents, just the indexes of what boxes --

THE COURT: I understand the words on the document don't mean anything, you just want what the document says it is.

THE WITNESS: Correct. It would show -- if there was a difference in the indexes that it returned it would indicate that there was some indexes not pulled and thus some boxes not pulled which may have documents in them if there was a difference between the alternate search query and the queries that were run.

THE COURT: But you don't know that at this time and you won't know it unless you run your system, let's call it that?

THE WITNESS: Correct.

THE COURT: Let's say you wanted to run it, how long would it take to run what we're talking about here? Is it confined to a certain area? Mr. Coenen, where are you? Mr. Coenen, are you suggesting a certain area of inquiry?

MR. COENEN: I think what happened in this other

case, it also involved Mr. Berteau, and our firm was, basically, they came up with some collaborative search terms and they ran them in connection --

THE COURT: I'm talking about what they're going to be searching for.

MR. COENEN: Certain area, Your Honor? I think we're looking for the things that are responsive to our discovery, but if you want specific key terms I'd have to think about that.

THE COURT: No. That's not what I'm asking. I'm asking, what do you want him to research with his system? All the records of UNOCAL and Chevron and Carrollton and all these people?

MR. COENEN: I think that's the only way we can figure out if they didn't produce certain things, Your Honor. As I mentioned earlier we were getting --

THE COURT: Let's assume they want you to do all that, how long is it going to take to do that?

THE WITNESS: Generally, if it was prepared beforehand, we would show up with a list of search queries that we wanted to run, and it would take maybe a day to pull the index, the number of indexes and the indexes that were pulled, and assuming that they retained the indexes they pulled the first time, or these search queries could be re-run, it should be a day or so. It's really not that difficult to run search queries. The pulling of the boxes is a different matter, but seeing if there is disparity between indexes returned from this and indexes returned from a more effective search would be relatively quickly.

THE COURT: In other words, you would look at

what has been discovered and then you would look at yours and compare them to see what hasn't been given over; is that the idea?

THE WITNESS: Correct. Exactly.

MR. BERTEAU: Your Honor, can I be heard on that a little bit? I hear their nice, fuzzy description of what has transpired in the prior cases, obviously, I disagree with the magnitude of the undertaking, the amount of attorney hours that got poured into it in the end. Your Honor, I hope you will let me go on the record here because this is a matter of grave concern to us.

This is our proprietary database. We don't want anybody looking in our index unless there is something that we've done wrong. I suggest to you that there isn't any proof in this as to anything we've done wrong.

THE COURT: Go ahead. I'm just trying to get it in my head.

MR. BERTEAU: I want to put another idea in your head if I could. I don't think we should have to look at any boxes that have been looked at before. I don't think we should have to re-do any work that's been done before. I would hope that the Court would give some credence and some credibility to the work that we've done in this case over the six years that it's been litigating and not just say -- because once this process starts I guarantee it's going to blow up unless we figure out some really tight restrictions that at least honor the work that we've done, Judge. I hope that what you'll find that what we're talking about here is a sanction without a wrongdoing, and

you'll find that it is inappropriate to grant them the relief that they're asking under the law.

THE COURT: I haven't ruled.

MR. BERTEAU: Okay. I understand. I just want to make sure.

MR. COENEN: For the record, we're not asking them to go back and do work they've already done. We're looking for documents that were not produced, obviously.

THE COURT: But you can't do that unless you go back and run the whole deal.

MR. COENEN: I think we can have an attorney present that's familiar with the documents that have been produced. That's how it happened in the Conoco and BP case. So if you see things that are returned that had already been produced, then we can pull that stuff out. That stuff that hasn't been produced, the attorney would be familiar with that and we could focus on those.

THE COURT: You guys do this all the time. Y'all all know one another. I'm new to this game, and I'm looking at it from a judge's standpoint, trying to understand it, and also from a juror's standpoint and trying to understand it.

MR. COENEN: I understand, Your Honor. And -- No.

THE COURT: So if my questions seem somewhat naive, I apologize. I'm just trying to get -- I hate to use this term. I'm not going to use it. I'm trying to understand it.

MR. BERTEAU: That's fine, Your Honor. I hope we can develop -- there's more to this case than just search terms, as I think our opposition makes clear.

THE COURT: Roll on.

EXAMINATION BY MR. BERTEAU:

Q. Mr. Roux, were you asked to apply your IT expertise, if I may call it that, to design search terms for the plaintiffs to run in response to UNOCAL's discovery in this case?

A. No.

Q. Have you even seen that discovery from UNOCAL to the plaintiffs?

A. No. Again, the only documents I reviewed are the search term list to provide an opinion as to its sufficiency and effectiveness as well as the deposition transcript from the individual.

Q. So you haven't spoken to anyone who works for the Vermilion Parish School Board other than the Carmouche firm; is that right?

A. That's correct.

Q. You don't have any information about any servers being maintained by the Vermilion Parish School Board; is that right?

A. That's correct.

Q. You don't have any idea about any7 electronic databases that may be maintained by the Vermilion Parish School Board?

A. That is correct.

Q. And had no role in the searches of any of those databases?

A. That's correct.

THE COURT: Have you seen the indexes that have been produced to date?

MR. BERTEAU: Your Honor, I was going to move to that right now, as a matter of fact, if I could. You're talking about the Court's order last time for them to produced their search terms?

MR. COENEN: He's talking about a different issue.

MR. BERTEAU: I'm not sure what you're asking about, Judge.

MR. COENEN: Your Honor, if I can maybe add some clarity to that. We don't have access to their actual indexes that were returned --

THE COURT: Start again. I missed the first thing you said.

MR. COENEN: We don't have access to the actual indexes that were produced. We just have access to the search terms. We don't have the indexes that were returned for each one of those searches that was generated.

MR. BERTEAU: Your Honor, I'm going to be candid, some of those exist, some of them don't. This goes way back in time, this even goes back in time before the onset of search terms. A lot of it -- as Mr. Wiebelt has testified in his deposition, an electronic result is generated, maybe it's saved and maybe it's not. It may be a learning process, like what Mr. Roux has testified about where you refine it and you keep looking. That's the danger of forming an assumption about the neatness of these kinds of projects. They're just not a neat project. They span years, a lot of people get involved. But we'll roll forward, with the Court's permission.

EXAMINATION BY MR. BERTEAU:

Q. Mr. Roux, do you have any information about the search -- any search work, any document review or production work done by UNOCAL in this case prior to the onset of search terms?

A. Again, the extent of my involvement has been to review

this search list for sufficiency and effectiveness and to review the deposition of the gentleman you just mentioned.

Q. Mr. Roux, do you have any knowledge about when the rule governing electronically stored information was adopted by the federal court system?

A. 2006.

Q. And it's true, is it not, that this case was filed in 2004?

A. I don't have any knowledge to that, again.

Q. No, I can see why you wouldn't. At the risk of stating the obvious, Mr. Roux, isn't it possible to find documents without using search terms?

A. From the database?

Q. Not from a database. Just documents.

A. I suppose you could wander around a warehouse aimlessly looking for them, but --

Q. Well, let's explore that a little bit. If instead of aimlessly wandering around a warehouse I talked to somebody who knows where the documents are stored and they take me to a room and say, here's the documents relating to East White field and I in look those boxes, I've conducted a documents search without using search terms; haven't I?

A. You could say that you've engaged in a search of employees, and that that particular employee doesn't necessarily contain all the information about where all the documents are stored. You would only be able to say that you've searched the documents that particular individual was aware of, or the subset of the total documents that individual was aware of.

Q. So you would find it, in your opinion, unlikely that the individual would know where all the documents are, which is -- he could know, couldn't he? It's possible?

A. Anything is possible.

Q. And if I talked to ten individuals, I increased the likelihood that I've captured that many more documents, right?

A. Not necessarily, because depending on where the individuals are situated in the company they may have overlapping knowledge or they may have knowledge of different areas of the total universe of documents, if that makes sense. If you asked everyone, for example, in Houston where their documents were stored, you'd get a lot of information about Houston documents, hypothetically; but if you didn't ask someone from other document storage locations or they had interactions with them, I would imagine you wouldn't be getting the full picture. But, again, that's not really my area of expertise, about personnel matters and physical document warehousing.

Q. Mr. Roux, did you have any role in the preparation of questions for Mr. Wiebelt, UNOCAL's 1442 --

A. Again, I reviewed his deposition after I reviewed these, and I was only brought in to look at this a couple of weeks ago. So, no.

Q. So if you had any questions about the Times database, how it operated, how searches were run in the Times database, you would have had an opportunity to explore those issues had you been called upon to participate in Mr. Wiebelt's deposition, correct?

A. Yes.

Q. Mr. Roux, in a -- strike that.

MR. BERTEAU: Your Honor, I'm going to mark as Exhibit 2, Defense 2, some documents given to me at the deposition of Kurt Soileau. Mr. Soileau is the Section 16 Supervisor for the School Board. I'll mark as Defendant 3 some excerpts from his deposition.

MR. COENEN: I'm just going to object to the general relevancy of different database.

THE COURT: Let me hear the question. It's overruled at this time.

MR. BERTEAU: I want to ask Mr. Roux if he's ever seen these documents, Your Honor.

EXAMINATION BY MR. BERTEAU:

Q. Mr. Roux, I have three sets of documents, index volume 23, 24, and 25, which I'll represent to you were given to me by the School Board. Just ask you if you've ever seen these documents before (handing)?

A. (The witness examines the documents.) No.

MR. BERTEAU: I will offer and introduce excerpts from Mr. Soileau's deposition as Defense 3. I just want to make sure it goes into the record.

THE COURT: Is there any objection to that? How are you getting this in? He never saw this before.

MR. COENEN: Your Honor, I'm going to object. I don't think this is even from the Soileau deposition. This looks like it's from a court transcript.

MR. BERTEAU: I'm sorry. This is an excerpt from the court order, Your Honor. Forgive me.

MR. COENEN: I would object to the introduction of the indices.

MR. BERTEAU: Let me explain what I am doing. First of all, of course, the Court ordered and the School Board agreed to produce their search terms. And I realize that's a parallel matter, but I think it's tied to this. And the reason that I think it's tied to this is because Article 1461 says when a party reconsiders that electronically stored information has not been produced in compliance with the request they can come to you for relief. What I suggest to you is that a party considering brings into play what exactly

has the role of search terms been in this case, including what has the role of search terms been in the plaintiff's own response to discovery.

The reason that I bring it to you is that I think it's a double standard. It's opportunistic, and it's hypocritical, and the sole purpose that it's brought before you is to try to get this trial pushed off and create a big ruckus. They could have done this at any time during the six-year period of time that this case has been pending. They chose to wait until the last minute. I think it goes to motive, Your Honor, and I think 1461 brings motive into play by its choice of words. So all I want to do is bring into the record what they have represented to us in response to the Court order as being the search terms that they used and make sure that that goes into the record, and the fact that Mr. Roux has never played any part in the development or use of any search terms in the plaintiffs responding to our discovery.

MR. COENEN: Your Honor, nothing that he just said makes that an admissible exhibit in connection with this witness. He can talk about that in his argument in opposition to this motion, but it has nothing to do with this cross-examination.

THE COURT: I agree. The objection is sustained.

MR. BERTEAU: Your Honor, I would suggest to you in response that it's a deposition of a party and an exhibit to that deposition. Please note my objection.

THE COURT: It's denied. Mr. Roux, let me ask you this, you saw the search terms that they used to do the search? You saw what was used, apparently?

THE WITNESS: The UNOCAL document?

THE COURT: Yes.

THE WITNESS: Yes.

THE COURT: When was that done? Was there a date on when those search terms were developed?

THE WITNESS: These seem to have a series of dates.

THE COURT: You tell me.

THE WITNESS: It has a series of dates, it seems, on when they were conducted over the course of several years. And in some cases done twice.

THE COURT: What are some of the dates? When were those search terms developed and when were some of the runs?

THE WITNESS: It seems to span from late 2005 through 2010 in April and May.

MR. BERTEAU: Could I have one second, Your Honor, to talk to my co-counsel?

That's all I have, Your Honor. If the Court will hear argument, obviously, I'll give you some of that.

THE COURT: You have some more questions?

MR. COENEN: I just have one redirect, that's it.

THE COURT: Go ahead.

EXAMINATION BY MR. COENEN:

Q. Mr. Roux, in order to determine the effectiveness of search terms, the search terms that are identified in front of you as Plaintiff's Exhibit 1, you would not just need to see those search terms but also the indexes that are returned from those searches, right?

A. That's correct. I would have to see both the indexes that these returned as well as indexes from revised search terms to compare the results.

MR. COENEN: Thank you, Mr. Roux. Your Honor, I

would offer, file, and introduce Plaintiff's Exhibit 1, which is a list of the search terms. May I approach?

THE COURT: It is admitted. You can step down. Thank you, Mr. Roux.

MR. COENEN: Your Honor, if you want to hear some argument, I can provide that now.

THE COURT: Well, where are the indexes?

MR. COENEN: The indexes, I believe, Your Honor, would have to be answered by Chevron. We don't have them. They're actually -- when they run a search they get a list of hits as to what boxes, I would assume, or what relevant -- possibly relevant documents are returned, and then they would, I guess, pull from those based on whatever search terms they actually ran. So we don't have those lists.

THE COURT: What did you give them? What did y'all get on that -- out of this thing?

MR. BERTEAU: Your Honor, I gave them the search terms. They never asked for the search spreadsheets.

MR. COENEN: I mean, they wouldn't even give us the search terms until Your Honor ordered it a couple of weeks ago. It would be very helpful to look at the search spreadsheets. We don't have that either. That's something I think they would've maintained -- if they're going to maintained the search terms themselves --

THE COURT: When did you get that document? The one that you just admitted.

MR. COENEN: That's the one that Your Honor ordered them to produce two weeks ago in court.

THE COURT: Go ahead, Mr. Coenen.

MR. COENEN: Your Honor, to get back to the search terms, these were run over the course of several years; however, the only time we realized that we were maybe missing some documents was whenever first I deposed Mr. Wiebelt who was the records custodian for UNOCAL. I asked him, Mr. Wiebelt, have you produced all the documents? Yes, I believe we have, all of them. Then I get -- a month later I'm starting to get more documents rolling in, so that's when I become suspicious that maybe I don't have everything. That is why we're here today, Your Honor, that's what led to the filing of this motion to compel.

You heard Mr. Roux testify, he said that they didn't run searches for VPSB as an acronym, which is an extremely important acronym in this particular case. They didn't run searches for East White Lake by itself. They ran East White Lake with other search terms, but they didn't run it by itself. The more search terms you add on to a search, the more narrow the search actually becomes, because it will only reveal -- if you ran East White Lake plus UNOCAL, for example, you're only going to get documents that come back that are indexed with both East White Lake and UNOCAL. It's very possible that things could be indexed just under East White Lake, or just under pit, or just under VPSB. That wasn't done in this case.

He also indicated, Mr. Roux did, that there are numerous problems and gaps with the searches themselves, the way they've used plural words or didn't use plurals in certain situations. He said those types of qualifiers can lead to big gaps in your

searches. But the best way for him to figure out whether or not these documents are missing is to actually run a search on their system, see what it returns and compare it with what the documents we have now. We can do that just like we did in the *Harris* case in the Conoco and BP databases, looking at real-time feedback, real-time searches and saying it looks like we don't have this particular document for our system, for this case.

Your Honor, that's what we submit. We submit that that's the only way to do it effectively.

MR. BERTEAU: Your Honor, is there any chance I can get a cup of water?

THE COURT: Yes, come on up here and you can pour yourself some right here.

MR. BERTEAU: Thank you.

THE COURT: By the way, I've told y'all before, if you want to bring in water or soft drink or anything. No margaritas or anything like that.

MR. BERTEAU: Your Honor, what Mr. Coenen is proposing in bland, innocuous terms that he presents to you is, in fact, a radical new rule of law that the case law does not support. What he's saying in essence is, if we think, just if we think we don't have everything, we get to check, we get to go behind you and check everything that you've done.

Now, starting with the law, article 1461, which discusses electronically stored information and the circumstances under which a party may be allowed access to a hard drive specifically requires not only the party considers themselves, but there's good cause to require it. It's a sanction. It's a sanction

because it's going to impose more man-hours, more cost, more review, and there's not a single piece of paper that's been proved to be missing in this case. There's not a single witness --

THE COURT: How can you prove what you don't know?

MR. BERTEAU: But that's the kicker, Judge, is we're having to prove the completeness of what we did, and it sounds like the 100 percent accuracy of what we did, instead of Mr. Coenen having to prove what we didn't do, which is where the burden of proof normally lies when you have a discovery dispute and a party is applying to this court for sanctions.

Now, This focus on search terms does a huge disservice to the fact that the case is clear and there's no dispute. We were looking for documents in this case that the attorneys who represented UNOCAL before Kean Miller was involved were looking for documents in this case, and we produced documents in this case that were done the old fashioned way. We went into a warehouse; we looked at the East White Lake documents. We went through them; we found the stuff that was relevant and we produced it, just like they did in the old days. And I don't think anybody would suggest that if one of our employees subsequently turned out to be a habitual drunk that they get to go back and redo the whole discovery process.

I mean, the integrity of a process, the integrity of production has to be based, first of all, just the way the Code is structured has to based on proof that there's something wrong, something tangible wrong,

actual proof that we did something wrong, that we have something and that we withheld it.

THE COURT: What about the main allegation, that the acronym for the School Board wasn't included in that query?

MR. BERTEAU: Your Honor, first of all, you asked me to produce the search terms that we have a record of, and I gave you the search terms that we have a record of. Does that represent all the search terms? I don't know. There could have been others. This has been a long case. It's had a lot of fingers on the case.

Now, you talk about the results. I don't have all the results. I'm going to tell you that I don't have all the results of the search terms over the five-year term that the case has run. I don't think there's ever been -- as a prudent person I would like that to have happen, but it just didn't happen in this case.

Now, I've got some of them, but the bottom line is, if "VPSB" was not used as a search term, which, in fact, it may have been; nevertheless, what is the likelihood that the other 14, 15 pages of search terms that you see before you would've resulted in the production of the box, the same box, all the boxes. There was 22,000 pages that were produced here.

I note to you, Judge -- I know you --

THE COURT: Let me see that, what was offered with the witness. Go ahead. You can go on.

MR. BERTEAU: This is a folder level index. What that means is each entry that you see on a spreadsheet, whether it's sent to you by computer or

in hard copy, whatever, is a folder, so any hit that you get -- in other words, the first one on there, "pits," I think that's the first one, if you get a hit on there, it's going to be on a folder. What that means is you're going to go pull that whole box and everything that's in that box is going to be subject to be reviewed.

So just the process itself when you have a folder hit, you capture everything in that box with the other folders, all the only other folders, which are naturally, probably, often, you know, even hopefully going to also contain documents that relate to the field. So it's not a scientific process, Judge. It's not a -- it's a process that many steps along the way there's possibilities of human error: somebody makes a mistake in spelling when they put it in, somebody makes a mistake when they go to pull a box off a shelf and they put it in a place where it shouldn't go and somebody doesn't see it and it comes back later. I don't know. It's a possibility of error.

If you give a list of search terms run by anybody, anybody in history of the world since the onset of electronic databases and you hire someone like Mr. Roux or anybody like -- any IT guy, you know they're going to come in here and say, aha, you should have run an extra "and," there should have been another "or," there's a universe of stuff you didn't capture.

Judge, isn't it appropriate, when the burden of proof is on Mr. Coenen to require some kind of showing that something is missing before we allow this kind of second guessing as to the adequacy of a six-year

process and require us to go back through it all over again.

THE COURT: What came out -- look on this exhibit 1, Plaintiff Exhibit 1.

MR. BERTEAU: I'm looking at it, Judge.

THE COURT: Look on the last page.

MR. BERTEAU: You mean the 73010?

THE COURT: 73010, "test" and "VPSB."

MR. BERTEAU: I don't know what specific box was generated as a result of running these --

THE COURT: Was a box generated?

MR. BERTEAU: Boxes were generated -- you'll notice that a bunch of searches were run at the same time.

THE COURT: Yes, I'm seeing that.

MR. BERTEAU: Okay. Boxes were generated as a result of those searches.

THE COURT: Those are the boxes I saw last time in court?

MR. COENEN: No, Your Honor. That's a totally different issue. VPSB -- the search term "VPSB" and "test," and we can have Mr. Roux testify about this, but if it's a Boolean type search it would only reveal documents that have been indexed under "VPSB" and also include the occurrence of the word "test." So it would be unlikely that it would return everything having to do with VPSB.

MR. BERTEAU: All I know is boxes were generated as a result of those searches. We looked through the boxes; we produced some documents. It's not a perfect process.

THE COURT: 730 "sample" or "sampling" and

"VPSB."

MR. COENEN: Again, that would be the same issue, Your Honor. It will only come back if you have those words in there, plus "VPSB."

MR. BERTEAU: That's what they were looking for, Judge.

MR. COENEN: Not "VPSB" by itself, Your Honor.

MR. BERTEAU: They're looking for samples, so we looked for samples.

MR. COENEN: Well, we're looking for a lot more than that. I think Your Honor would agree.

Your Honor, I think what -- the problem here is that by use of this database he says it can favor us because we can keep coming back and say your search wasn't good enough, but it more greatly favors Chevron who can run a bad list of search terms and maybe in the hopes they'll miss some documents. That's actually a possibility. It benefits them greatly to have these databases and only produce things that are returned by the databases.

I think you heard Mr. Berteau just say -- I think I heard him say this, that there are probably additional search terms we don't have and there are probably additional documents if you keep looking for these things. That's what we -- we need certainty. We need to know that we have every single thing concerning this case, and up to this point I don't have that assurance. I just don't have it, Your Honor.

MR. BERTEAU: Your Honor, the Code is not shaped around whether or not Mr. Coenen feels reassured. The Code is shaped around whether or not there is specific

evidence of wrongdoing or deficiency in the production process. Now, why over the course of six years would there be not a hint of an inadequacy of search terms run or any suggested alternatives. This issue didn't come up until a month before the original schedule for trial, just like it was in *Harris*. Yeah, I'll make the comparison with *Harris* to that extent. It's a naked procedural maneuver to cast doubt on something that has not been shown through any tangible evidence to be deficient.

Judge -- let me finish -- the suggestion -- we did a lot of work in this case, Judge. I don't think there's any doubt about that. I think Mr. Roux even concedes that in his affidavit. He said I've got a problem even if they did all the work, even if they did a thorough review. Don't make us do anything again that we've already done. And there's no indication of anything that we have not done except submit ourselves to the tender mercies of Mr. Coenen and his expert. And that's a sanction, Judge, and we -- there is no basis for a sanction in this case.

THE COURT: Has Mr. Roux prepared a suggested list of terms, Mr. Coenen?

MR. COENEN: He hasn't done it yet, Your Honor, but we can if need be, quickly.

Your Honor, while we're at it, if we're going to talk about *Harris,* I think it's important for the Court to know, was it done a month before trial? Because -- yes, because we did not know about documents missing until that time. In fact, when we did get into their database and ran searches we were getting documents pertaining to the very field that

involved that case the week before trial, and BP searches continue even to this day. We're getting documents that were never produced by BP. I mean, so this is --

THE COURT: You're talking about the other case now.

MR. COENEN: That's the other case. I just wanted to give Your Honor an example. If Your Honor would like Mr. Roux to prepare some search terms, we can certainly do that.

MR. BERTEAU: And all I'm saying, Judge, in the absence of wrongdoing don't reopen discovery in this case. You're doing it on their whim. They haven't run the first search term in their database. They don't care about search terms. The only reason they care about it -- they could have done a million different things to create this issue from being brought to you now by way of foresight if they truly cared about search terms and the adequacy of production. This should not be an issue being brought to your attention 30 days before trial with this kind of IT second-guessing.

MR. COENEN: Your Honor, the only tool I have -- if I take a deposition of a records custodian, I'm entitled to rely on what that records custodian tells, and if he tells me he's given me everything, that's fine, I don't really have much other place to go. But if he then turns over and dumps a bunch of documents on me that should have been covered by my original searches, then I have a problem.

THE COURT: Mr. Roux, come on back up here again.

(Whereupon, Mr. Roux retakes the witness stand.)

THE COURT: You're still under oath, Mr. Roux.

EXAMINATION BY THE COURT:

Q. Mr. Roux, if you wanted to do what you suggested to improve the quality of this search, what would you do?

A. I would have to take a list of pertinent terms fir the case, go through the list of searches that they did craft, swap out terms that are either --

Q. You've looked at that already, haven't you?

A. I have the list of what they were searching for. I know there's some terms that they were not searched by themselves, that were not searched in combination with each other.

Q. So you wouldn't duplicate what's already been done?

A. No. I would take searches that would be comparable to those to see if a different -- more results were returned or a different subset of results were returned, if that makes sense. It would compare directly the size of the result set based on changing the terms or making some of them more broad so that they're not limited, in the case of the acronym, seeing how many were returned by just the acronym versus the acronym conjoined with the test, I think, and to see how much of a difference, if any, there is between the two.

Q. How long would it take you to prepare those terms?

A. It should only take a day or so to come up with a list of terms.

Q. And then a day to run it?

A. That should -- that would be a good estimate.

Q. And you wouldn't duplicate anything that's been done before?

A. Not in terms of the document review. The indexes may be returned in a duplicative fashion, but that could be culled out by comparing the two index result sets. So if you get A, B

and C from the search term they originally ran, but you get A, B, C and D from the revised term you can cull out the A, B, and C and you're just left with the D.

THE COURT: Natalie, let me ask you one thing.

(DISCUSSION OFF THE RECORD)

THE COURT: If you want to ask him any questions -- any other questions, you may.

MR. BERTEAU: No, Your Honor. I'd just like to address specifically --

MR. COENEN: I have no other questions, Your Honor.

THE COURT: Thank you, Mr. Roux. You can step down.

MR. BERTEAU: I don't -- I want to follow the right order.

MR. COENEN: Your Honor, we are willing to make this as least invasive as possible. We're willing to follow whatever Your Honor says, obviously, but as Mr. Roux said, it can be done fairly quickly. We can have lawyers involved to the extent necessary to make sure no attorney-client privilege information is being divulged, just like we did in the other case. We can -- however Your Honor prefers it be structured. We're open to suggestion.

MR. BERTEAU: Your Honor, if the Court is inclined, over our strenuous objections, to grant them relief, what I would ask would be let us run the searches and let us generate the results.

THE COURT: With their terms?

MR. BERTEAU: With their terms. And we'll look it over and make sure there's nothing privileged on the result itself, and then we'll see where we go from

there.

MR. COENEN: The problem with that is, Your Honor, is there's no real-time feedback for our guy, our IT expert, just like he was saying he needed in order to carry this out efficiently. What Mr. Berteau is talking about doing is going to take a lot longer, I think. We're going to have to get the search terms back and you don't have any real-time exposure to this database by doing that. All you can do is look at what they give back to us and say, well, now we know we've got this. How do we know if they culled something else. I mean, it's not effective to do it that way.

MR. BERTEAU: Your Honor, the Court is considering imposing a sanction against us. All I ask is that you at least let us try to honor our privacy interest in this database. They've got the IT expert. He's presumably been thinking about this for eons. Let him give us the search terms, and we'll run the search terms and give him the spreadsheet. I mean, he ought to be able to -- he looked at what we've run. He ought to be able to come up with something worthwhile. Him standing there looking at our computer --

THE COURT: Why is that going to take any longer than any other method?

MR. COENEN: I think we could have Mr. Roux testify to that, too. He's testified on the stand that doing it while he's there is far more efficient than waiting until they run a search -- list of search terms, coming back to us. We've got to look at them and see, well, okay -- we don't know what they got

back in their indexes. We don't know what was returned by the search terms that were giving them. I think it's a difficult situation to do it that way.

MR. BERTEAU: Your Honor, if they are truly -- if he has truly examined these search terms it should not be difficult for him to come up with search terms that we haven't run yet that would be suitable for whatever their purposes are.

MR. COENEN: We don't know what their indexes are going to come back with. What if you run "VPSB" and your index comes back VPSB 1, 2, 3, or something like that and we don't know it -- we've got to explore what that document is. It's important to know on a real-time basis what is going on with this database. They can't just run things and tell us, well, we found this; we didn't find that. It's a lot harder for us to sit back and figure it out.

THE COURT: That's what they've been doing, huh?

MR. COENEN: Well, that's what we counted on them to do, but I think in this situation we know the original terms didn't work.

THE COURT: You don't know it. You believe that the quality of the search terms didn't reveal everything that you believe you should have gotten. And you don't know that until you run it with the search terms.

All right. Let's see, today is Wednesday. The Court is going to order that a limited list of search terms be developed and that it be run out of the presence of any of the experts. It will be done with only the defendant's lawyers and any other representatives they want present. And this is to be

accomplished by Monday at four o'clock. All right. So be finished by then.

MR. COENEN: The search will actually be finished by Monday is what Your Honor is suggesting?

THE COURT: That's correct. You get your terms together and get them to them and so that they can finish it up by Monday.

MR. BERTEAU: And part of the order would be the terms are not to include anything we've already run?

THE COURT: That's correct.

MR. BERTEAU: Note our objection for the record, Your Honor.

THE COURT: Nothing that's duplicative. The objection is noted and overruled.

MR. COENEN: Thank you, Judge.

THE COURT: Let's go to three, Plaintiff's Motion for Leave to File Fifth Supplemental and Amending Petition for Damages.

MR. MARCELLO: Your Honor, before the Court takes this up, I'd like to indicate to the Court our position as to what -- what our position is in the case at this time. I have prepared a motion and could not file it until the Court granted the suspensive appeal. I'll give a copy of it to the defendants now. I don't expect them to respond immediately to it.

THE COURT: What are you filing?

MR. MARCELLO: It's a motion and incorporated memorandum to stay and continue the trial based upon jurisdictional grounds, that the suspensive appeal causes the Court to lose jurisdiction. It involves matters particular to this particular case because of a prior Third Circuit ruling in this case which stated

Vita

The author was born and raised in the Greater New Orleans area. He attended Brother Martin High School, and later obtained his Bachelor's degree in Computer Science from the University of New Orleans in 2007. After returning for graduate school, he earned his Master's degree in Computer Science with concentration in Information Assurance from the University of New Orleans in 2008 while also pursuing research in Bioinformatics as a research assistant to Dr. Winters-Hilt. He continued on to enroll in the PhD program at UNO, and is also concurrently pursuing a JD (expected May 2013) at the Tulane University Law School with particular interests in intellectual property, international, and comparative law.